DOCUMENT RESUME

ED 318 056 CS 507 127

AUTHOR	Pisoni, David B.; And Others
TITLE	Research on Speech Perception. Progress Report No. 12.
INSTITUTION	Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY	Air Force Armstrong Aerospace Medical Research Lab, Wright-Patterson AFB, OH.; National Institutes of Health (DHHS), Bethesda, Md.; National Science Foundation, Washington, D.C.
PUB DATE	86
CONTRACT	AF-F-33615-83-K-0501
GRANT	BNS-83-05387; NS-07134-08; NS-12179-10
NOTE	457p.; For other reports in this series, see CS 507 123-129.
PUB TYPE	Reports - Research/Technical (143) -- Collected Works - General (020) -- Information Analyses (070)
EDRS PRICE	MF01/PC19 Plus Postage.
DESCRIPTORS	*Acoustic Phonetics; Auditory Discrimination; *Auditory Perception; Communication Research; Computer Software Development; Infants; *Language Processing; Language Research; Linguistics; Speech; *Speech Synthesizers
IDENTIFIERS	Indiana University Bloomington; *Speech Perception; Speech Research; Theory Development

ABSTRACT

Summarizing research activities in 1986, this is the twelfth annual report of research on speech perception, analysis, synthesis, and recognition conducted in the Speech Research Laboratory of the Department of Psychology at Indiana University. The report contains the following 23 articles: "Comprehension of Digitally Encoded Natural Speech Using a Sentence Verification Task (SVT): A First Report" (D. B. Pisoni and M. J. Dedina); "Comprehension of Natural and Synthetic Speech: II. Effects of Predictability on Verification of Sentences Controlled for Intelligibility" (D. B. Pisoni and others); "Perceptual Learning of Synthetic Speech Produced by Rule" (S. L. Greenspan and others); "Trading Relations, Acoustic Cue Integration, and Context Effects in Speech Perception" (D. B. Pisoni and P. A. Luce); "Using Template Pattern Structure Information to Improve Speech Recognition Performance" (M. Yuchtman and H. C. Nusbaum); "On Word-Initial Voicing: Converging Sources of Evidence in Phonologically Disordered Speech" (J. A. Gierut and D. A. Dinnsen); "On the Assessment of Productive Phonological Knowledge" (J. A. Gierut); "Generative Phonology and Error Pattern Analyses: Empirical Claims and Differences" (J. A. Gierut); "Effects of Talker Uncertainty on Auditory Word Recognition: A First Report" (J. W. Mullenix and D. B. Pisoni); "Effects of Stress and Final-Consonant Voicing on Vowel Production: Articulatory and Acoustic Analyses" (V. Summers); "Preference Judgments Comparing Different Synthetic Voices" (J. S. Logan and D. B. Pisoni); "Auditory Perception of Complex Sounds: Some Comparisons of Speech vs. Nonspeech Signals" (D. B. Pisoni); "Perceptual Attention in Monitoring Natural and Synthetic Speech" (H. C. Nusbaum and others); "Intelligibility of Phoneme Specific Sentences Using Three Text-to-Speech Systems and a Natural Speech Control" (J. S. Logan and D. B. Pisoni); "PRONOUNCE: A Program for

Pronunciation by Analogy" (M. J. Dedina and H. C. Nusbaum); "The Role of the Lexicon in Speech Perception" (D. B. Pisoni and others); "The Role of Structural Constraints in Auditory Word Recognition" (H. C. Nusbaum and D. B. Pisoni); "A Brief Overview of Speech Synthesis and Recognition Technologies" (D. B. Pisoni); "Developing Methods for Assessing the Performance of Speech Synthesis and Recognition Systems" (D. B. Pisoni and H. C. Nusbaum); "Recognition Performance of Six Isolated Utterance Speech Recognition Systems" (H. C. Nusbaum and others); "Human Factors Issues for the Next Generation of Speech Recognition Systems" (H. C. Nusbaum and D. B. Pisoni); "Using Speech as an Index of Alcohol Intoxication" (C. S. Martin and M. Yuchtman); "Effects of Wholistic versus Dimensional Training on Learning to Identify Spectographic Displays of Speech" (B. G. Greene); and "Testing the Performance of Isolated Utterance Speech Recognition Devices" (H. C. Nusbaum and others). (SR)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED318056

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 12
(1986)

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

Supported by

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. NS-12179-10

National Institutes of Health
Training Grant No. NS-07134-08

National Science Foundation
Research Grant No. BNS-83-05387

and

U.S. Air Force
Armstrong Aerospace Medical Research Laboratory (AFSC)
Contract No. AF-F-33615-83-K-0501

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. B. PISONI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CS507127

BEST COPY AVAILABLE

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 12

(1986)

David B. Pisoni, Ph.D.

Principal Investigator

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

Research Supported by:

Department of Health and Human Services
U. S. Public Health Service

National Institutes of Health
Research Grant No. NS-12179-10

National Institutes of Health
Training Grant No. NS-07134-08

National Science Foundation
Research Grant No. BNS 83-05387

and

U. S. Air Force
Armstrong Aerospace Medical Research Laboratory (AFSC)
Contract No. AF-F-33615-83-K-0501

Table of Contents

Introduction iii

I. Extended Manuscripts 1

Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report; David B. Pisoni and Michael J. Dedina 3

Comprehension of natural and synthetic speech: II. Effects of predictability on verification of sentences controlled for intelligibility; David B. Pisoni, Laura M. Manous, and Michael J. Dedina 19

Perceptual learning of synthetic speech produced by rule; Steven L. Greenspan, Howard C. Nusbaum, and David B. Pisoni . . . 43

Trading relations, acoustic cue integration, and context effects in speech perception; David B. Pisoni and Paul A. Luce 87

Using template pattern structure information to improve speech recognition performance; Moshe Yuchtman and Howard C. Nusbaum . . 107

On word-initial voicing: Converging sources of evidence in phonologically disordered speech; Judith A. Gierut and Daniel A. Dinnsen 125

On the assessment of productive phonological knowledge; Judith A. Gierut 151

Generative phonology and error pattern analyses: Empirical claims and differences; Judith A. Gierut 175

Effects of talker uncertainty on auditory word recognition: A first report; John W. Mullennix and David B. Pisoni 205

Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses; Van Summers 223

Preference judgements comparing different synthetic voices; John S. Logan and David B. Pisoni 263

II. Short Reports and Work in Progress 291

Auditory perception of complex sounds: Some comparisons of speech vs. nonspeech signals; David B. Pisoni 293

Perceptual attention in monitoring natural and synthetic speech; Howard C. Nusbaum, Steven L. Greenspan, and David B. Pisoni . . . 307

Intelligibility of phoneme specific sentences using three text-to-speech systems and a natural speech control; John S. Logan and David B. Pisoni 319

PRONOUNCE: A program for pronunciation by analogy; Michael J. Dedina and Howard C. Nusbaum 335

The role of the lexicon in speech perception; David B. Pisoni, Paul A. Luce, and Howard C. Nusbaum 349

The role of structural constraints in auditory word recognition; Howard C. Nusbaum and David B. Pisoni 361

A brief overview of speech synthesis and recognition technologies; David B. Pisoni 369

Developing methods for assessing the performance of speech synthesis and recognition systems; David B. Pisoni and Howard C. Nusbaum 379

Recognition performance of six isolated utterance speech recognition systems; Howard C. Nusbaum, C. Noah Davis, David B. Pisoni and Ella Davis 389

Human factors issues for the next generation of speech recognition systems; Howard C. Nusbaum and David B. Pisoni 403

Using speech as an index of alcohol intoxication; Christopher S. Martin and Moshe Yuchtman 413

Effects of wholistic versus dimensional training on learning to identify spectrographic displays of speech; Beth G. Greene . . . 427

III. Instrumentation and Software Development 439

Testing the performance of isolated utterance speech recognition devices; Howard C. Nusbaum, Christopher K. Davis, David B. Pisoni, and Ella K. Davis 441

IV. Publications 457

V. SRL Laboratory Staff and Personnel 461

INTRODUCTION

This is the twelfth annual report summarizing the research activities on speech perception, analysis, synthesis, and recognition carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various research activities over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software support when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating our own research.

We are distributing reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of speech processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis, synthesis, and recognition and, therefore, we would be grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
USA
(812) 335-1155

Copies of this report are being sent primarily to libraries and specific research institutions rather than individual scientists. Because of the rising costs of publication and printing, it is not possible to provide multiple copies of this report to people at the same institution or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address.

The information contained in the report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.

I. EXTENDED MANUSCRIPTS

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Comprehension of Digitally Encoded Natural Speech Using a Sentence Verification Task (SVT): A First Report*

David B. Pisoni and Michael J. Dedina

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*This research was supported in part by NIH Research Grant NS-12179 and in part by Air Force Contract No. AF-F-33615-83-K-0501 from the Armstrong Aerospace Medical Research Laboratory (AFSC) to Indiana University in Bloomington. We thank Dr. Thomas J. Moore of Wright-Patterson AFB for his help in processing the stimulus materials used in this study and his continuing interest and support.

Abstract

A sentence verification task (SVT) was used to measure comprehension of short sentences of digitally encoded natural speech. Three different vocoders were tested: (1) 16 kbps continuously variable slope delta modulation (CVSD), (2) 9.6 kbps time domain harmonic scaling subband coding (TDHS/SBC), and (3) 2.4 kbps linear predictive coding (LPC). Subjects listened to sentences produced by one of these vocoders and were required to determine if each sentence was "True" or "False". Subjects also transcribed the sentence after each trial. Three dependent measures were obtained: (1) verification accuracy, (2) verification response latency, and (3) sentence transcription accuracy. The following rank ordering was found across the three measures: CVSD was comprehended the most accurately, followed by TDHS/SBC, and then LPC. The difference in comprehension performance in the SVT between the highest-ranked and lowest-ranked vocoders was statistically reliable and robust for all three dependent measures. The present findings demonstrate that the SVT is an extremely sensitive technique to measure comprehension processes that are not indexed well by traditional tests of segmental intelligibility such as the modified rhyme test (MRT), which is concerned with phoneme and feature perception in isolated monosyllablic words. When listeners are asked to "understand" the linguistic content of a message and to execute an appropriate response, the quality of the initial acoustic-phonetic information in the speech signal appears to play an important role in controlling both the speed and accuracy of the response.

9

Comprehension of Digitally Encoded Natural Speech

Using a Sentence Verification Task (SVT): A First Report

This paper reports the results of a study that examined the comprehension of short naturally-spoken sentences using a sentence verification task (SVT). In a recent study in our laboratory, Manous, Pisoni, Dedina and Nusbaum (1985) found that sentence verification error rates, response latencies and transcription scores provided very sensitive measures of listeners' performance that could be used to index processing load in the comprehension of synthetic speech produced by rule. The comprehension scores in this task for ten synthetic voices also correlated very highly with measures of segmental intelligibility obtained in earlier experiments using the Modified Rhyme Test. The results suggested that differences in the early stages of processing the acoustic-phonetic input appear to cascade up the speech processing system to influence comprehension as well as recognition processes. Large and statistically reliable differences were obtained in latencies among systems even when the sentences were accurately transcribed thus ensuring that the sentences were correctly encoded on input and not misperceived.

In the present study, we were interested in learning if these differences in perceptual processes were due to properties of synthetic speech per se or whether they may reflect a more general property of speech signals that have been processed so as to reduce the information content of the linguistic message by coding and bandwidth reduction techniques. It is a well-known observation that the speech signal is extremely redundant and that the human listener can tolerate massive distortions and degradation in the signal characteristics without noticeable loss in intelligibility or comprehension of the linguistic message. Would differences similar to those found with synthetic speech also be observed with digitally encoded speech?

To study this problem, we examined the perceptual consequences of using three different algorithms for digitally encoding natural speech produced by both a male and female talker. The data reported by Manous et al. (1985) on the comprehension of synthetic speech using the SVT procedure demonstrated extremely reliable differences among synthetic voices even though the differences in some of the scores were relatively small. We hoped that this same procedure could be used to reveal differences in the perception of natural speech that was processed using digital encoding techniques. The three digital signal processing techniques used in this study were: (1) continuously variable slope delta modulation (CVSD); (2) time domain harmonic scaling subband coding (TDHS/SBC) and (3) linear predictive coding (LPC). The CVSD has a data rate of 16 kbps, the TDHS/SBC has a data rate of 9.6 kbps and the LPC-10, the DoD government standard, has a bit rate of 2.4 kbps. The natural speech for this study was produced by one male talker and one female talker. Both were native speakers of English and came from the midwestern region of the United States.

Given the recent findings of Nixon, Anderson and Moore (1985) on the perception of natural speech, digitally encoded speech, and synthetic speech in noise using the modified rhyme test (MRT), we were interested in determining the consequences of using the three different speech encoding algorithms on comprehension performance in the SVT. Nixon et al. (1985) reported fairly small differences in the MRT scores obtained in noise for these same three digital speech coders over the range of S/N ratios that were studied. The TDHS/SBC speech coder was the worst of the three systems even

10

though the overall difference was only about ten percent. The LPC was the best, with the CVSD falling somewhere in the middle. As expected, unprocessed natural speech showed substantially higher MRT scores at all S/N ratios tested compared to the digitally processed speech and the synthetically produced speech. Of particular interest in the Nixon et. al. (1985) study, however, was the unexpected finding that high-quality synthetic speech generated automatically by rule was very close to natural speech over the range of S/N ratios studied. Overall, MRT scores for this particular text-to-speech synthesizer were higher than any of the systems tested including the three digital speech coders using natural speech and the two other text-to-speech systems.

Although all current speech synthesis systems produce speech that is mechanical and unnatural sounding, the segmental intelligibility of several of these systems appears to be extremely high, approaching that obtained with natural speech (see Pisoni, Nusbaum & Greene, 1985; Greene, Logan, & Pisoni, 1986). Based on early research on speech synthesis at Haskins laboratories in the early 1950's, there is some reason to believe that naturalness and segmental intelligibility are orthogonal perceptual dimensions and that even highly mechanical-sounding speech may still be highly intelligible and may remain so even under adverse listening conditions. Indeed, the very earliest studies of speech synthesis by rule using the Haskins Pattern Playback produced highly intelligible although admittedly quite unnatural sounding speech because of the monotone fundamental frequency used to synthesize the harmonic series (Cooper, Liberman & Borst, 1951; Cooper, Liberman, Borst & Gerstman, 1952).

The purpose of the present investigation was to measure the comprehension of short sentences of digitally encoded natural speech using the sentence verification task developed recently in our laboratory by Manous et al. (1985). Although Nixon et. al (1985) reported small differences in MRT scores among the three digital speech coders, we hoped that a more sensitive test employing response latencies and transcription scores would reveal a more robust pattern of differences that could be compared with the earlier results of Manous et al. (1985) obtained using synthetic speech produced by rule. Moreover, we were interested in determining if differences at the acoustic-phonetic level would affect high-level processes associated with comprehension of the linguistic message.

Method

Subjects. The subjects were undergraduate students enrolled in an introductory psychology course at Indiana University. All were native speakers of English who reported no history of a speech or hearing disorder at the time of testing. The subjects fulfilled a course requirement in introductory psychology by participating in this study. Between 13 and 16 subjects participated in each of the 6 conditions of the experiment, for a total of 90 subjects. None of the subjects had any previous experience in listening to synthetic speech before the present study.

Stimuli. The test items were sixty short English sentences originally developed by Dr. Harry Levitt of CUNY for testing a visual speech display system for hearing impaired subjects (Weiss, Levitt, & Halprin, 1983). All sentences were screened in our laboratory by two independent judges to eliminate items that were potentially ambiguous or inappropriate for an auditory comprehension task. The final materials used in the experiment were 30 three-word and 30 six-word sentences; half of the sentences at each length were "True" and half were "False." The false sentences could not be falsified until the subject heard the last word in the sentence. Thus, for both the true and false sentences, subjects had to listen to the entire sentence to respond correctly. Four additional sentences were chosen as practice items to familiarize the subjects with the task and materials.

Six tokens of each of the 60 test sentences were produced for the experiment. Each item was recorded on audio tape by a male talker (PAL) and a female talker (JCL). The tape was then sent to the Air Force Aerospace Medical Research Laboratory at Wright-Patterson Air Force Base for processing. The sentences were processed by the three digital encoding techniques mentioned above, then recorded on audio tape and sent back to our laboratory. The sentences were then low-pass filtered at 4.8 kHz, and digitized at 10 kHz using a 12-bit A/D converter and edited into individual stimulus files using an interactive waveform editor.

Procedure. Subjects were run in groups ranging in size from three to five listeners. Each subject sat at a booth equipped with high-quality matched and calibrated headphones (Telephonics TDH-39) and a two-button response box that was interfaced to a PDP-11/34 computer. At the beginning of each session, the experimenter read aloud the instructions to the subjects while they read a printed version in front of them. Subjects were told that they would hear a short sentence on each trial and that their task was to determine if the sentence was "True" or "False." Then subjects received four practice trials to familiarize themselves with the task and with the sound quality of the particular type of speech used in each condition. Following the practice trials, 60 experimental trials were presented. The entire experiment lasted about a half hour.

Test sentences were presented to the subjects over headphones. All stimulus materials were output using a PDP-11/34 computer via a 12-bit D/A converter. On each trial, subjects first heard a sentence and then made a forced-choice True/False response by pressing one of the appropriately labelled buttons on the two-button response box. Subjects were instructed to respond as quickly and accurately as possible when making their responses. After they had entered their response on each trial, subjects were required to transcribe the entire sentence by writing down exactly what they heard on a separate response sheet. When all of the subjects finished transcribing the sentence, the next trial began. The experimenter remained in the experimental room during the course of the experiment to ensure that subjects were responding appropriately. The trials were paced to the slowest subject in each group. Response latencies were measured from the onset of each sentence to the subject's response, using special purpose computer-controlled routines. The duration of each sentence was then subtracted from the measured response latency to provide a true measure of response time from the end of the sentence that is not contaminated by differences in stimulus length.

Results

The data were analyzed using three different dependent measures: (1) sentence verfication error rate, (2) verification response latency, and (3) sentence transcription error rate. Separate ANOVAs were carried out for each of the three dependent measures to determine the effects of vocoder, voice gender, and sentence length on subjects' performance. For each dependent measure, "True" and "False" responses were analyzed separately. In these analyses, vocoder and voice gender were between-subjects factors and sentence length was a within-subjects factor.

Sentence Verification Accuracy. The error rates for each voice at each sentence length are displayed in Figures 1a and 1b for "True" and "False" responses, respectively. ANOVAs revealed a significant main effect of vocoder for both "True" and "False" responses ($F = 32.93$, $p < .0001$ and $F = 49.17$, $p < .0001$), a main effect of gender for "True" responses ($F = 5.33$, $p < .0235$), and a main effect of length for "False" responses ($F = 7.27$, $p < .0085$). In addition, the ANOVA for "True" responses showed a length by vocoder interaction ($F = 3.16$, $p < .0461$), and a length by vocoder by gender interaction ($F = 3.62$, $p < .0310$). The ANOVA for "False" responses showed a main effect of length ($F = 7.27$, $p < .0095$), and interactions between vocoder and gender ($F = 17.04$, $p < .0001$), and length and vocoder ($F = 9.37$, $p < .0002$). Newman-Keuls post-hoc analyses were carried out to determine which vocoders differed significantly from the others. The analyses were done separately for each voice gender. For the male voice, LPC had a significantly higher error rate than TDHS/SBC and CVSD. TDHS/SBC and CVSD did not differ from each other. For the female voice, CVSD had a lower error rate than LPC and TDHS/SBC, but LPC and TDHS/SBC did not differ.

Insert Figures 1a and 1b about here

Sentence Verification Latencies. Response latencies were analyzed only for sentences that had been both verified correctly *and* transcribed correctly verbatim. The reason for using only these trials was to insure that the sentences were encoded correctly on input and that any observed differences were not due to misperceptions at the time of perceptual encoding. Thus, any differences in the pattern of latencies would have to be due to processes related to comprehension of the message. Verification latencies are shown in Figures 2a and 2b for "True" and "False" responses, respectively. Figures 3a and 3b show the proportion of the total responses in each voice condition which were included in calculating the verification latencies.

Insert Figures 2a, 2b, 3a and 3b about here

For true sentences, an ANOVA revealed a significant main effect of vocoder ($F = 53.13$, $p < .0001$), and an interaction between length and gender ($F = 7.05$, $p < .0095$). For false sentences, significant main effects were found for vocoder ($F = 51.43$, $p < .0001$) and sentence length ($F = 32.52$, $p < .0001$), and interactions were found between voice and gender ($F = 4.27$, $p < .0172$), and length and vocoder ($F = 13.69$, $p < .0001$). Newman-Keuls post-hoc analyses revealed the same pattern for response latencies as we found for error rates in the previous analysis. For both the male and female voices, CVSD was responded to significantly faster than LPC. With the male voice, TDHS/SBC differed from LPC but not from CVSD. However, with the female voice, TDHS/SBC differed from CVSD but not from LPC.

Sentence Transcription Accuracy. Sentence transcription data were hand-scored for the absolute number of errors in correctly transcribing the key words in each sentence. A word was scored as a correct response if it was

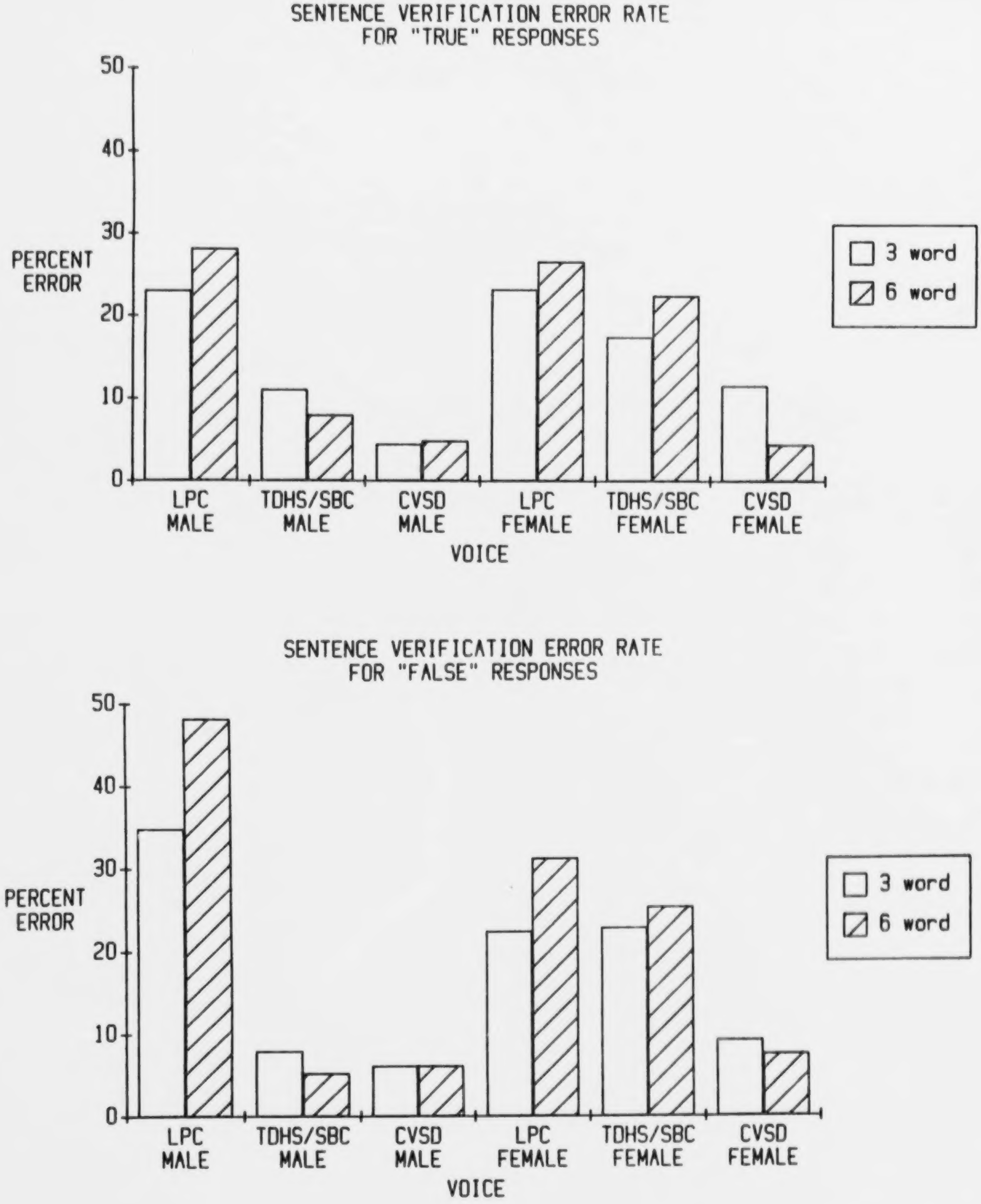


Figure 1. Sentence verification error rates for "True" responses (top panel) and "False" responses (bottom panel) for the six different voices. Three-word sentences are shown with open bars, six-word sentences are shown with striped bars.

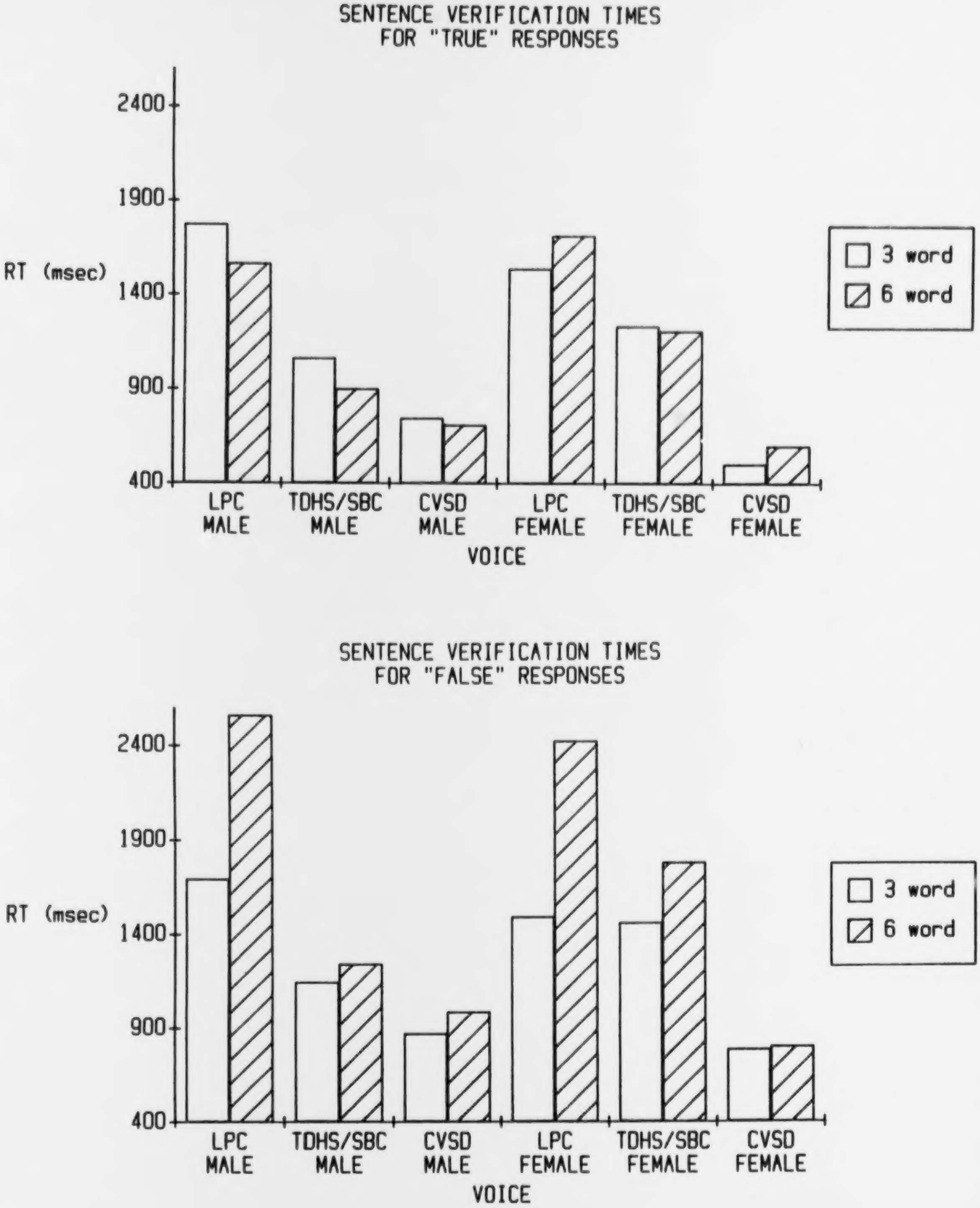
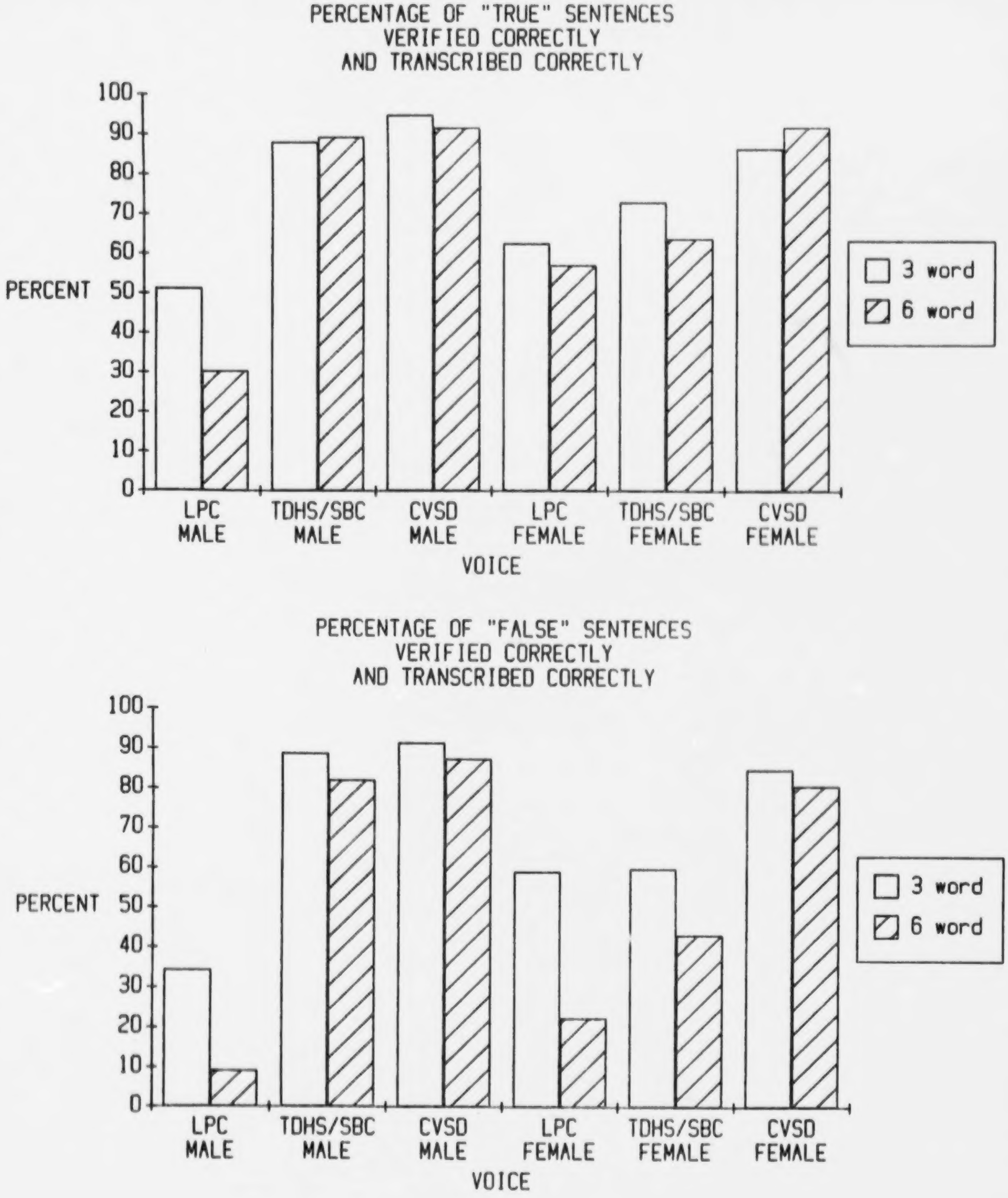


Figure 2. Average sentence verification latencies (in msec) for "True" responses (top panel) and "False" responses (bottom panel) for the six voices. These latencies are based on only the trials in which the subject responded correctly and also transcribed the sentence correctly. Three-word sentences are shown with open bars, six-word sentences are shown with striped bars.



PERCENTAGE OF "TRUE" SENTENCES VERIFIED CORRECTLY AND TRANSCRIBED CORRECTLY

PERCENTAGE OF "FALSE" SENTENCES VERIFIED CORRECTLY AND TRANSCRIBED CORRECTLY

Figure 3. Percentage of trials in which subjects responded correctly and transcribed the sentence correctly. "True" sentences are shown in the top panel; "False" sentences are shown in the bottom panel. Three-word sentences are shown with open bars, six-word sentences are shown with striped bars.

an exact phonemic match of the corresponding key word. Spelling mistakes were ignored. The number of errors was then used to compute an error percentage for each subject on true and false sentences of each length. These data are shown in Figures 4a and 4b.

Insert Figures 4a and 4b about here

ANOVAs on these transcription data revealed a main effect of vocoder for both true and false sentences (F = 216.75, p < .0001; F = 434.49, p < .0001). For true sentences, significant interactions of vocoder with gender (F = 77.21, p < .0001) and length with gender (F = 5.29, p < .0239) were observed. In addition, a three-way interaction between length, vocoder, and gender was found for true sentences (F = 5.06, p < .0084). A significant main effect of length (F = 91.31, p < .0001), and interactions of vocoder with gender (F = 61.16, p < .0001), and length with vocoder (F = 36.63, p < .0001) were also observed for the false sentences. Post-hoc tests showed a similar pattern for transcription errors as we found in our analyses of the other two dependent measures. Sentences produced by CVSD were transcribed more accurately than those produced by LPC in all conditions. With the male voice, TDHS/SBC differed from LPC but not CVSD for both "True" and "False" responses. With the female voice, TDHS/SBC differed from CVSD but not LPC for "True" responses. Using the sentence transcription scores, however, all three vocoders differed significantly with the female voice for "False" responses. CVSD was transcribed most accurately, and LPC was transcribed least accurately. This pattern of results differs slightly from the findings obtained using the other two dependent measures.

Discussion

The results of the present study demonstrate reliable and extremely robust differences in comprehension of short sentences processed by three digital vocoders. The overall ranking on all three measures -- verification accuracy, verification latency, and sentence transcription accuracy parallels the data rate of the vocoders under examination. The worst system, the LPC-10, had a data rate of 2.4 kbps whereas the best system, the CVSD, had a data rate of 16 kbps. These two systems also differ in the kind of encoding algorithm used to process speech. The LPC is primarily a technique involving analyses of speech signals in the spectral domain whereas the CVSD involves analyses in the time domain. We also observed a reliable and very consistent interaction between vocoder and talker. The TDHS/SBC was consistently worse for the female talker than the male talker across all three dependent measures.

Setting aside the gross differences among the encoding algorithms used in these vocoders, the results of the present study are of interest in connection with the findings reported recently by Nixon et al. (1985) on the intelligibility of digitally encoded speech in noise and the earlier study from our laboratory by Manous et al. (1985) on the comprehension of synthetic speech produced by rule. Using the MRT procedure, Nixon et al. (1985) found very small differences among the three digital vocoders used here. No statistical analyses of the differences were reported in their paper so we can only infer from our own visual examination of their figures that the

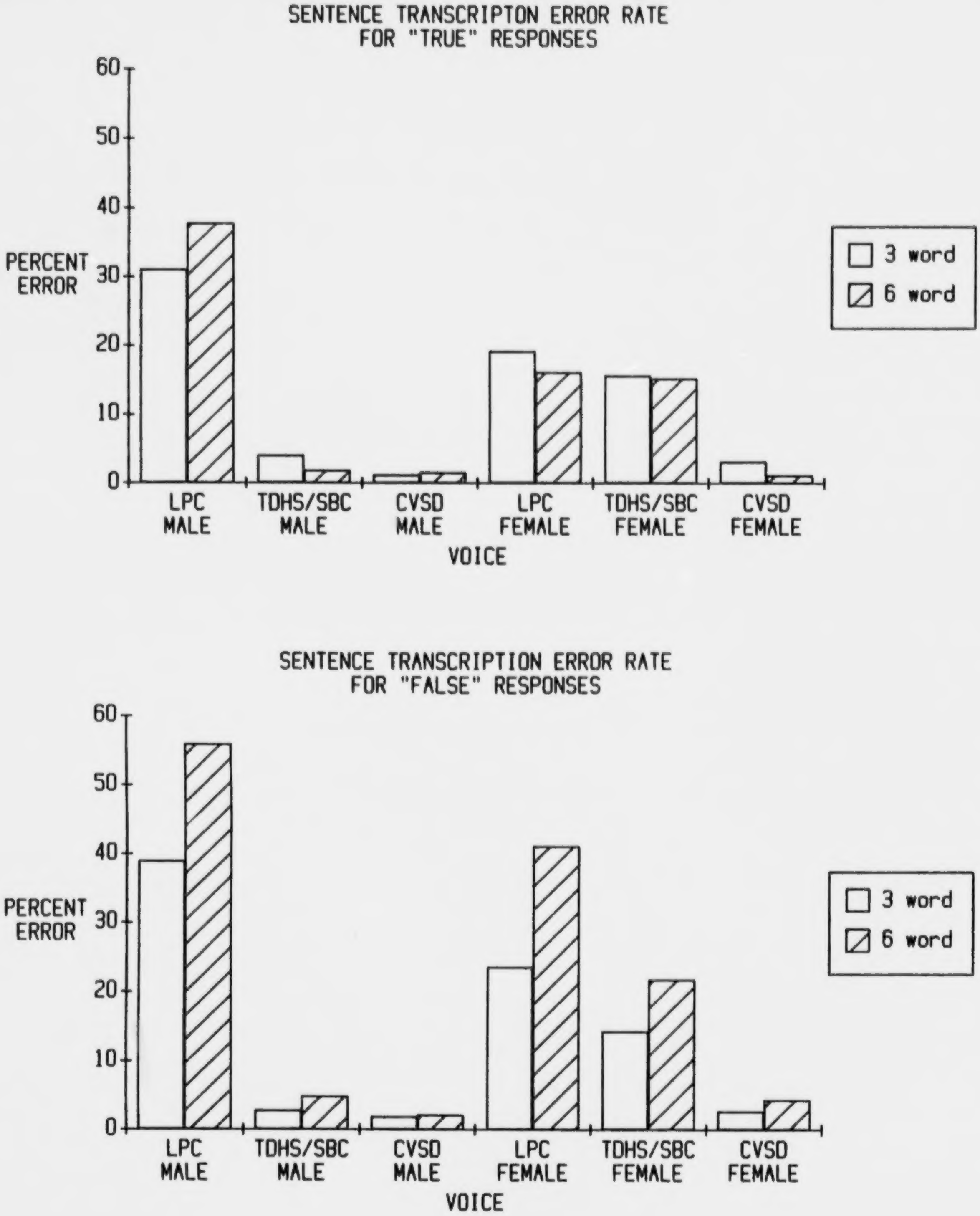


Figure 4. Percentage of errors obtained in the sentence transcription task for "True" responses (top panel) and "False" responses (bottom panel). Open bars represent data for three-word sentences, striped bars represent data for six-word sentences.

differences were probably not statistically reliable. In contrast, using the same three vocoders, we obtained large and consistent differences in the comprehension of short sentences. Thus, the present procedure using the SVT task appears to be much more sensitive to quality differences among vocoders than traditional forced-choice measures of segmental intelligibility using isolated monosyllabic words.

These comprehension findings which were obtained for both verification accuracy and response latencies are more impressive when one considers that the differences were obtained on test trials in which the observer correctly transcribed the sentence. Thus, differences in segmental intelligibility due to misperceiving of the input signal could be effectively ruled out as an explanation of the observed differences among the systems. In other words, the test items were correctly encoded on input, and therefore the differences we obtained must reflect some process or processes above the level of phoneme recognition. Whatever the nature of these differences turns out to be, it is clear that the speech processing system is somehow registering, and then passing on to higher processing levels, properties of the speech signal related to its initial segmental intelligibility. Although a test sentence may be correctly transcribed in terms of its lexical content, there appears to be a cost associated with encoding a degraded signal, and this cost is cascaded up the processing system. While these quality differences may not be reliably indexed by traditional intelligibility measures such as the MRT or DRT, the present procedure is able to separate out differences among these three vocoders in a reliable and robust manner. These differences in comprehension performance reflect the underlying perceptual and cognitive processes involved in extracting the linguistic message from the speech signal and responding appropriately to the meaning of the linguistic information encoded in the signal rather than simply responding to the linguistic form of the message. The results of this study demonstrate important differences among three widely used digital vocoders in a task that requires more than simple recognition and subsequent transcription of the message.

The findings obtained in the present study also extend the earlier research reported recently by Manous et al. (1985) on the comprehension of synthetic speech produced by rule. Reliable differences in the same three dependent measures were obtained for natural speech and several types of synthetic speech. The SVT task was shown to be extremely sensitive to differences among different systems generating synthetic speech by rule. Like the present results, the findings suggested that differences in the processing of the acoustic-phonetic input affect verification accuracy and latency and, therefore, appear to cascade up the processing system to affect higher-level operations typically associated with comprehension. For purposes of comparison, we present the overall error rates for both studies. Figure 5 shows the verification error rates found by Manous et. al for five text-to-speech systems and natural speech, alongside data from the present study using three digital vocoders. The data in this figure are averaged across gender and sentence length. Performance on TDHS/SBC and CVSD are roughly equivalent to the two lowest quality synthesizers, while the error rate for LPC is almost twice as high as for the lowest quality synthesizer. These results are consistent with the data reported by Nixon et al. (1985) showing higher levels of performance with synthetic speech produced by rule compared to natural speech processed by several different digital encoding algorithms.

Insert Figure 5 about here

In addition to observing reliable differences among the three vocoders, a number of other findings emerged from our analyses of the comprehension scores. Some of these differences reflect processes that would ordinarily be related to the linguistic analysis of the message rather than the perceptual analysis of the signal. For example, for false sentences we found an extremely reliable effect of sentence length. The six word false sentences consistently produced more verification errors and were responded to more slowly than the three word sentences. This finding suggests that the differences among the processing algorithms are not restricted to only analyses of the segmental and lexical content of the signal. Rather, processing operations typically associated with comprehension and linguistic analysis of the sentence structure are affected as well even when the sentences are correctly perceived and encoded. Further studies of these processing differences would no doubt be very worthwhile to determine the nature of the differences and to locate the specific operations within the language processing system.

We also observed several complex interactions between vocoder and gender suggesting that some algorithms are not well suited for encoding of female speech. These findings are not at all surprising. For example, the error rates and latencies for the 9.6 kbps TDHS/SBC algorithm were consistently higher for the female talker than the male talker reflecting, in part, the difficulty encountered in dealing with the higher fundamental frequencies of female talkers and the greater spacing of the harmonics of the source spectrum. Some findings on differences among talkers using LPC techniques have been reported in the literature by Kahn and Garst (1983). Data on talker variability using other digital encoding techniques have also been described by Smith (1979). Thus, there is good reason to expect that talker differences should also emerge in tasks such as the present one which appears to be even more sensitive than MRT and DRT scores to factors affecting intelligibility and comprehension.

Much of the research on speech communication over the last forty years has been concerned with behavioral measures of segmental intelligibility using materials such as PB lists, MRT or DRT tests (Voiers, 1977). These techniques are easy to administer and score and they provide extremely reliable data that reflects the quality and sufficiency of the segmental acoustic-phonetic properties of the speech signal. These traditional methods were not designed to study the cognitive processes that mediate spoken language comprehension. Taken together with the earlier study by Manous et al. (1985), we believe the present findings using the SVT task may provide an extremely powerful method to study the underlying perceptual and cognitive processes that mediate between the speech waveform and understanding the content of the linguistic message. The SVT task is able to discriminate large and consistent differences between several digital encoding algorithms that are not revealed very well using the conventional MRT or DRT measures. Our findings suggest that segmental intelligibility is only one component, although obviously a very important one, in the comprehension process. Additional findings using techniques such as the SVT should prove extremely useful in studying the interface between speech perception and spoken language comprehension and in decomposing the processing stages and operations. The fact that spoken

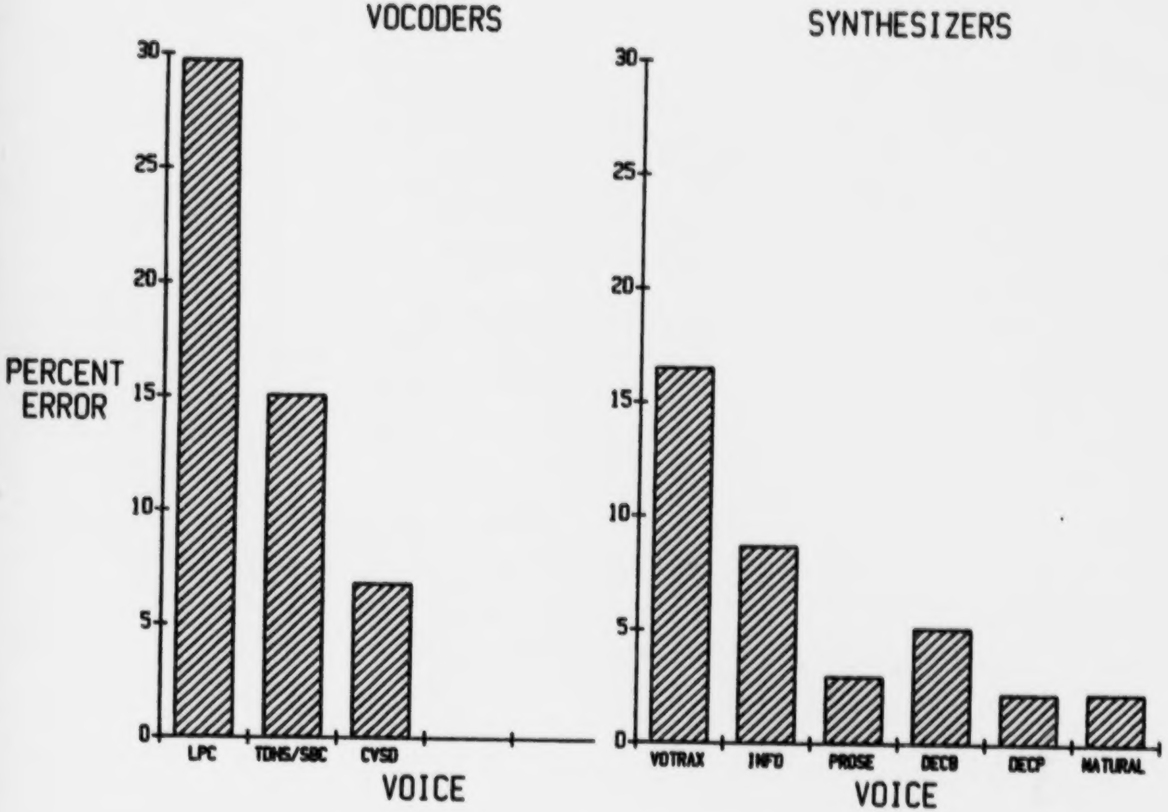


Figure 5. Sentence verification error rates for the three vocoders examined in the present study (left panel), and for the five text-to-speech systems and natural speech examined by Manous, Pisoni, Dedina, and Nusbaum (1985) (right panel).

language is so redundant and that our speech processing system is so robust often obscures the basic operations we wish to study and understand. We are encouraged, however, by the results obtained in the present study and the earlier findings of Manous et al. (1985). It may now be possible to begin to study comprehension processes in ways that will reveal the close interdependence between the acoustic-phonetic information in the speech signal and the rich and varied sources of knowledge and linguistic constraints that listeners have available to them as native speakers of the language.

References

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*, *37*, 318-325.

Cooper, F. S., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*, 597-606.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, and Computers, 18 (2)*, 100-104.

Kahn, M. & Garst, P. (1983). The effects of five voice characteristics on LPC quality. *Proceedings of 1983 IEEE International Conference on Acoustics, Speech and Signal Processing*, April, Boston, Massachusetts.

Manous, L. M., Pisoni, D. B., Dedina, M. J., & Nusbaum, H. C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. *Research on Speech Perception Progress Report No. 11*. Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Nixon, C. W, Anderson, T. R., & Moore, T. J. (1985). The perception of synthetic speech in noise. In R. Salvi, D. Henderson, R. P. Hamernik, & V. Coletti (Eds.), *Applied and Basic Aspects of Noise Induced Hearing Loss*. NY: Plenum, in press.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, *73*, 1665-1676

Smith, C. (1979). Talker variance and phonetic feature variance in diagnostic intelligibility scores for digital voice communications processors. *Proceedings of the 1979 IEEE International Conference on Acoustics, Speech and Signal Processing*, April, Washington, D.C.

Voiers, W. D. (1977). Diagnostic evaluation of speech intelligibility. In M. E. Hawley (Ed.), *Speech intelligibility and speaker recognition*. Stroudsburg, PA: Dowden-Hutchinson and Ross.

Weiss, M., Levitt, H., & Halprin, M. S. (1983). Reading speeds for visual displays. Paper presented at Acoustical Society Meeting, Cincinnati, Ohio, May, 1983.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Comprehension of Natural and Synthetic Speech: II. Effects of Predictability on the Verification of Sentences Controlled for Intelligibility*

David B. Pisoni, Laura M. Manous, and Michael J. Dedina

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This research was supported by NIH Research Grant NS-12179 to Indiana University in Bloomington. We thank Howard Nusbaum for his helpful comments and suggestions.

Abstract

A sentence verification task (SVT) was used to study the effects of sentence predictability on comprehension of natural speech and synthetic speech that was controlled for intelligibility. Sentences generated using synthetic speech were matched on intelligibility with natural speech using results obtained from a separate sentence transcription task. In the main experiment, the sentence verification task included true and false sentences that varied in predictability. Results showed differences in verification speed between natural and synthetic sentences, despite the fact that these materials were controlled for intelligibility. This finding suggests that the differences in perception and comprehension between natural and synthetic speech go beyond segmental intelligibility as measured by transcription accuracy. The observed differences in response times appear to be related to the cognitive processes involved in understanding and verifying the truth value of short sentences. Reliable effects of predictability on error rates and response latencies were also observed. High predictability sentences displayed lower error rates and faster response times than low predictability sentences. However, predictability did not have differential effects on the processing of synthetic speech as expected. The results demonstrate the need to develop new measures of sentence comprehension that can be used to study speech communication at processing levels above and beyond those indexed through transcription tasks or forced-choice intelligibility tests such as the MRT.

Comprehension of Natural and Synthetic Speech: II. Effects of Predictability on the Verification of Sentences Controlled for Intelligibility

Over the past six years, numerous studies on the perception of synthetic speech have been conducted in our laboratory at Indiana University (see Pisoni, 1982; Pisoni, Nusbaum and Greene, 1985; Nusbaum and Pisoni, 1985). The bulk of these studies have focused on measures of segmental intelligibility such as identification of isolated words and recognition of words in fluent sentences (e.g., House, Williams, Hecker and Kryter, 1965; Egan, 1948; Nye and Gaitenby, 1973). Results from these studies of phoneme and word perception have shown that synthetic speech is consistently less intelligible than natural speech (see Greene, Logan, and Pisoni, 1986). This was observed for a variety of synthesis systems ranging from very low-quality to extremely natural sounding speech.

Most perceptual studies dealing with segmental intelligibility have not addressed the issue of comprehension processes involved in understanding the linguistic content of the message. In tests of segmental intelligibility such as the ones we have carried out, subjects are not required to extract or compute the meanings of utterances in order to make appropriate responses. They can carry out the task based on their discrimination of the acoustic-phonetic properties of the speech alone without fully understanding what they are listening to, or making a response that is based on the comprehension of the message. Depending on the type of comprehension test employed, subjects must use other information to generate a correct response.

To date, relatively little work has been done to examine how listeners comprehend synthetic speech produced automatically by text-to-speech systems. Speech quality, or overall intelligibility of the input signal, is certainly an important factor involved in spoken language comprehension. Yet additional consideration must also be given to the contribution of higher sources of knowledge in "understanding" the message and responding appropriately to the truth-value of sentences.

The few studies that have been conducted to examine comprehension of natural and synthetic speech have produced equivocal results, making it difficult to draw any general conclusions about the comprehension process. In one early study, McHugh (1976) assessed comprehension of synthetic and natural speech using passages selected from a standardized reading comprehension test. Prosodic information was manipulated by presenting six different stress variations of the synthetic speech from the Votrax synthesizer along with a natural speech control condition. Subjects' performance showed no significant differences across the seven conditions. McHugh concluded that the test she used was too sensitive to individual differences in performance to reveal any difference between the various experimental versions of speech that were tested.

In studies carried out in our laboratory Pisoni (1979) and Pisoni and Hunnicut (1980) studied the comprehension of natural speech and synthetic speech produced by MITalk, a text-to-speech system developed at MIT (see Allen, 1981). Listening comprehension was compared to reading comprehension for identical passages, using multiple-choice questions taken from standardized reading comprehension tests. Pisoni and Hunnicutt's results demonstrated that naive listeners were able to comprehend passages of synthetic speech at levels comparable to subjects who either heard passages of natural speech or who read the passages and answered the same questions after presentation of each passage.

In another study using passages of connected speech, Jenkins and Franklin (1981) examined comprehension of natural speech and synthetic speech produced by a Votrax text-to-speech system, using a free recall task and a sentence dictation procedure. One group of subjects transcribed a passage presented one sentence at a time. Another group of subjects listened to the entire passage and then attempted to recall the information just presented, in a free recall format. Once again, the results showed little difference in performance between natural and synthetic speech. Apparently, the behavioral measures used to assess comprehension were too gross and insensitive to reveal differences between various types of speech.

More recently, Schwab, Nusbaum and Pisoni (1985) included listening comprehension passages and true-false questions along with other tests to study the effects of perceptual learning on the perception of synthetic speech. As in the previous studies, the comprehension task did not reveal any effects of training or any differences between natural and synthetic speech. These results were surprising because all of the other tests (e.g. identification of isolated words, recognition of words in fluent sentences) used to assess performance in this study showed significant effects of training on the perception of synthetic speech.

Following up on these earlier comprehension studies, Moody and Joost (1986) have recently examined listener comprehension rates for synthesized speech using DECtalk, digitized speech using 9.6 and 2.4 kbps LPC algorithms, and natural recorded speech. Passages and multiple-choice questions were selected from standardized verbal exams such as the SAT and GRE. Their results showed significant differences in question-answering performance for synthetic speech and 2.4 kbps LPC digitized speech compared to the natural speech. The difficulty of the passage affected comprehension rates for all passages, regardless of the type of speech signal used. However, Moody and Joost observed an unusual interaction between passage difficulty and speech type in their study. When subjects listened to more difficult information in some passages, differences in performance between the natural speech group and the synthetic speech group were not observed. However, when the comprehension materials were easy, significant differences between the natural and synthetic speech groups emerged.

It is not immediately obvious to us how one would account for these findings given the earlier study of Luce, Feustel, and Pisoni (1983) which showed increased error rates in serial recall when capacity demands of the task were increased. We do not know of any current theory of human information processing or language comprehension that would predict the results observed by Moody and Joost. If there is some relationship between comprehension difficulty and signal quality, then differences in performance among natural speech, synthetic speech, and digitally vocoded speech should emerge more robustly under experimental conditions in which there are greater capacity demands on the processes used in perception or comprehension. Resolution of this problem obviously awaits additional research on comprehension using long passages of connected speech that have been specifically designed to differ in comprehension difficulty. For the present, we simply wish to point out that research on comprehension of synthetic speech continues to yield equivocal results that are difficult to integrate with other findings reported in the literature. When a situation like this arises, it is often useful to examine some of the commonalities and differences in the experimental procedures that have been used in this research and consider alternative techniques that may be used to approach the same general problem.

When considered together, all of the previous studies on comprehension have a number of similarities. First, they all used post-perceptual measures to index differences in comprehension. It is well-known in the comprehension literature that post-perceptual measures are affected by a variety of subject strategies that rely on numerous sources of knowledge in addition to the linguistic information contained in the input signal. Second, these studies have all used multiple-choice or true-false question answering tasks or recall tasks which encourage subjects to exploit their real-world knowledge to solve the task. Finally, all of these studies have used accuracy measures to index processing load instead of response latencies. The consistent failure to find differences in perception between natural speech and several kinds of synthetic speech using these measures suggests the need for much more sensitive methods of measuring ongoing processing activities. One such method is the sentence verification task, which has been used extensively in previous psycholinguistic investigations of the language comprehension process.

Sentence verification has been used for many years to assess processing activities in studies on language perception and comprehension (see Clark and Clark, 1977). In one of the earliest studies using this procedure, Gough (1965, 1966) found that sentence verification time varied as a function of grammatical form. Reaction times were shorter for active as opposed to passive sentences, affirmative as opposed to negative sentences, and true as opposed to false sentences. Collins and Quillian (1969, 1970) and Conrad (1972) have used sentence verification to study the organization and retrieval of semantic knowledge about words in long-term memory (see Chang, 1986, for a recent review). Both studies used response time as a measure to infer the level of processing required to verify information contained in various types of sentences, such as "a canary is a bird" or "a canary has wings." More recently, Larkey and Danley (1983) used sentence verification to investigate the role of prosody in comprehension of digitally vocoded natural speech. They found that subjects were 48 msec slower in responding to sentences with a monotone pitch than to sentences with the original prosodic contour left intact.

In a recent study carried out in our laboratory, Manous, Pisoni, Dedina and Nusbaum (1985) used the sentence verification task to investigate differences in comprehension between natural speech and synthetic speech generated by five different text-to-speech systems. They found that response latencies to verify short sentences sentences increased as segmental intelligibility of the speech decreased. Specifically, the results yielded a reliable rank-ordering of the different voices in which level of performance corresponded to the quality of segmental information for each type of speech. That is, performance on the sentence verification task for the various voices followed the pattern observed in earlier standardized tests of segmental intelligibility (Greene, Logan, and Pisoni, 1986). These findings suggest that the early stages of the comprehension process depend primarily, if not exclusively, on segmental intelligibility. However, it is possible that other processes are also affected by the quality of the initial acoustic-phonetic input in the speech signal. Differences in the early stages of perceptual analysis of the input may cascade up the processing system and impact on other processes more closely related to comprehension.

The present study was designed to examine this issue more closely and to dissociate effects due to segmental intelligibility from those related to comprehension processes. By controlling the level of intelligibility of the speech, we hoped to assess the comprehension process more directly and to draw inferences about processing activities that were not confounded with initial differences in segmental intelligibility. To accomplish this, we matched high-quality synthetic speech produced by DECtalk with natural speech in terms of segmental intelligibility. We then used the sentence verification task to compare performance for these two types of stimulus materials using test sentences that varied in length and semantic predictability. If the differences in perception between natural speech and very high quality synthetic speech are not due only to segmental intelligibility, then we would expect to find differences in response times in a verification task even though the error rates were comparable. Such a finding would be an important demonstration that the perception and comprehension of synthetic speech differs in important ways from the processing of natural speech (Pisoni, 1982). Moreover, such a finding with stimulus materials controlled for segmental intelligibility would suggest that cognitive processes related to comprehension are also affected by the initial quality of the acoustic-phonetic input in the speech signal.

If synthetic speech speech is indeed more difficult to comprehend in some general sense than natural speech, this difference should be influenced by other factors that affect speech perception and spoken language comprehension. In order to investigate this hypothesis, we manipulated the predictability of the last word in the sentences. In low-predictability sentences, less contextual information is available from earlier context to facilitate the perceptual process. In this case, listeners must rely more heavily on the acoustic-phonetic input in these sentences, therefore drawing scarce processing resources away from high-level comprehension processes. Assuming that the human speech processing system has only limited processing capacity at its disposal, we expect that if synthetic speech is more difficult to understand than natural speech, a manipulation of predictability would have a larger effect on synthetic speech than on natural speech. In addition to manipulating sentence predictability, we also varied sentence length as a rough index of syntactic complexity. We expected to find interactions of these two variables with the voice manipulation. If sentence length is an index of syntactic complexity, we expected to find that long sentences would be more difficult to process than short sentences and that this effect would be reliably greater for synthetic speech than natural speech.

Method

Subjects. Subjects were either volunteers who were paid $3.50 for their participation in this study or introductory psychology students who participated to fulfill a course requirement. Subjects were drawn from the same general university population. An equal number of subjects from these two groups participated in each condition of the experiment. All were native speakers of English with no reported history of a speech or hearing disorder. None of the subjects had any extensive experience in listening to synthetic speech before the present experiment.

Stimuli. In the first phase of the experiment, test items were specifically developed to vary along the dimension of semantic predictability. These materials were generated by having subjects provide the final word to complete 100 3-word and 100 6-word sentence frames. Examples of these stimuli are given in Table 1. For half of the sentences of each length, subjects were instructed to create true sentences; for the other half, subjects were required to construct false sentences. Forty subjects participated in this phase of the experiment.

Insert Table 1 about here

The data from this task were scored in terms of response frequency for each item. The sentences were then categorized by response predictability. Sentences for which a high frequency of subjects gave the same response and for which there was only one response of high frequency were labelled "High Predictability." Out of 40 subjects, 25 or more had to respond with the same word in order for a sentence to be classified as "High Predictability." "Low Predictability" sentences were defined as sentences for which there was a unique response, that is, sentences for which only one subject gave that particular response. Examples of "High Predictability" and "Low Predictability" sentences are shown in Table 2.

Insert Table 2 about here

The second phase of the experiment involved intelligibility testing for the sentences that were generated using synthetic speech produced by rule. This phase was designed to match synthetically produced test sentences with natural test sentences for segmental intelligibility. The sets of "High Predictability" and "Low Predictability" sentences obtained in Phase 1 were recorded on audio tape using the DECtalk version 2.0 text-to-speech system. These sentences were then presented to subjects in a transcription task.

Twenty-four additional subjects listened to the sentences and transcribed each one as accurately as possible with paper and pencil. Transcriptions were scored for exact phonemic match to the original sentences. Spelling errors were ignored unless they affected the meaning of the sentence (e.g. 'medal' for 'metal').

Based on the data obtained in Phase 2, 40 true sentences and 40 false sentences were selected to be used in the main sentence verification task. For 77 of these sentences, there were no transcription errors; each of the remaining three sentences had only one transcription error. Half of the sentences selected were "High predictability" sentences; the other half were "Low predictability" sentences. In addition, half of the sentences of each type were 3-word sentences and half were 6-word sentences.

Additional tokens of each of the 80 test sentences were produced by a male talker (PAL). Both groups of test sentences, the synthetic speech and the natural speech materials, were low-pass filtered at 4.8 kHz, then digitized at 10 kHz using a 12-bit A/D converter and edited into individual stimulus files using a digital waveform editing program.

Procedure. Sixty subjects participated in the final phase of the experiment. Two to five subjects were run at a time in small groups. Each subject sat at a booth equipped with high-quality matched and calibrated headphones (Telephonics TDH-39) and a two-button response box. Stimulus presentation and response collection were controlled by a PDP 11/34 computer. At the beginning of each session, the experimenter read the instructions aloud to the subjects while they simultaneously read a printed version in front of

Table 1. Examples of Test Sentences Used for Predictability Norms

Three-Word Sentences:

Cotton is ____________________.

Birds can ____________________.

Six-Word Sentences:

Pots and pans are used for ____________________.

Most businessmen wear suits to ____________________.

Table 2. Examples of High and Low Predictability Sentences

HIGH PREDICTABILITY

3-word	-->	Giraffes are tall. (TRUE)
		Sandpaper is smooth. (FALSE)
6-word	-->	Pots and pans are used for cooking. (TRUE)
		Sunglasses are most useful at night. (FALSE)

LOW PREDICTABILITY

3-word	-->	Trees have greenery. (TRUE)
		Diamonds are rough. (FALSE)
6-word	-->	Most businessmen wear suits to lunch. (TRUE)
		Leap year comes every four minutes. (FALSE)

them. Subjects were told that they would hear one sentence on each trial and that their task was to determine if the sentence was "true" or "false". Each group of subjects heard only one type of speech; half of the subjects listened to natural speech, and half listened to synthetic speech. Sentence length and sentence predictability were within-subject factors.

Subjects received four practice trials to familiarize them with the task and with the sound quality of the voice used in that particular condition. Following the practice trials, 80 experimental trials were presented. Test sentences were presented to subjects over headphones, via a 12-bit D/A converter. On each trial, subjects first heard a sentence and then made a forced-choice true/false response by pressing one of the appropriately labelled buttons on a two-button response box. Subjects were instructed to respond as quickly and accurately as possible when making their true/false decisions. After entering their response, subjects were required to transcribe each sentence on a separate answer sheet. This task was included to ensure that subjects had correctly encoded the test sentences on input.

During the course of the experiment, the experimenter remained in the experimental room to ensure that subjects were responding appropriately. Test trials were paced to the slowest subject in each group. Response latencies were measured using computer-controlled routines from the onset of each sentence to the subject's response. The duration of each sentence was then subtracted from the measured response latency to provide a measure of response time that was not contaminated by differences in stimulus length.

Results

A few of our subjects seemed to ignore our request that they respond quickly in this experiment. In order to reduce the subsequent variability in our data, we omitted from our final analyses the subjects whose average response latencies were greater than two standard deviations from the mean. Using this criterion, three subjects were eliminated, two from the natural speech group, and one from the synthetic speech group. We also eliminated one additional subject from the synthetic speech group, so that the same number of subjects were omitted from each experimental condition. This last subject had the slowest mean response time of the remaining subjects in the synthetic speech group. Thus, the final analyses reported below were based on data collected from 56 subjects.

Sentence Transcription Scores. In order to confirm that we had, in fact, successfully controlled for the segmental intelligibility of the sentences across the two sets of stimulus materials used in the experiment, we analysed the effect of voice (natural vs. synthetic) on transcription accuracy. An analysis of variance on the transcription scores revealed no significant effect of voice for true sentences ($F(1,54) = .96$, N.S.) or for false sentences ($F(1,54) = .70$, N.S.). The data were then analyzed using the two other dependent measures: (1) sentence verification accuracy, and (2) response latency. Separate analyses were carried out for each dependent measure to assess the effects of the three experimental manipulations: (1) voice, (2) sentence length, and (3) sentence predictability. In carrying out these analyses, true and false sentences were analyzed separately. The experimental design included three main effects: voice was a between-subjects factor, whereas sentence length and sentence predictability were within-subjects factors. Unless otherwise noted, the significance levels below are reported for the $p < .01$ level of confidence.

Sentence Verification Accuracy. Figure 1 shows the verification error rates for true and false sentences. Overall, the error rates were quite low, demonstrating that subjects had little difficulty in understanding the sentences and carrying out the verification task with both natural and synthetic speech.

Insert Figure 1 about here

Inspection of the error rates shown in Figure 1 reveals several consistent effects of the experimental variables. For the true responses, displayed in the top panel of this figure, the observed error rates were consistently higher for low predictability sentences than high predictability sentences. This was found for both natural and synthetic speech and was observed at each of the two sentence lengths used in the study. Analysis of variance confirmed these observations for the effects of predictability on true sentences ($F(1,54) = 38.01$, $p < .001$). All other effects in analyses of the error rates for both true and false responses failed to reach significance.

Although there was a trend for the error rates to be slightly higher for the synthetic speech, the differences were not reliable in either analysis of the true or false sentences. This result is not surprising considering the procedures we used to match sentences on intelligibility before the main verification experiment was carried out. The absence of an effect of voice in the analysis of the verification error rates is also consistent with the analyses of the transcription data described earlier in which no differences were found in immediate recall between the natural and synthetic sentences. Thus, taken together, both sets of data--the transcription scores and the sentence verification error rates, suggest that subjects correctly encoded the stimulus materials at the time of input and that they comprehended the linguistic content and meaning of the sentences. Although a reliable effect of sentence predictability was observed for the true sentences, the absence of a main effect for voice combined with the absence of any interactions with the voice manipulation suggests that the differences in the perceptual encoding between the natural and synthetic stimuli were minimal at best. In short, the expected outcome for both of these measures was observed.

Verification Response Latency. Response latencies were analyzed only for sentences that had been both verified correctly and transcribed correctly. Figure 2 shows the mean response latencies for true and false sentences in each of the conditions of the design.

Insert Figure 2 about here

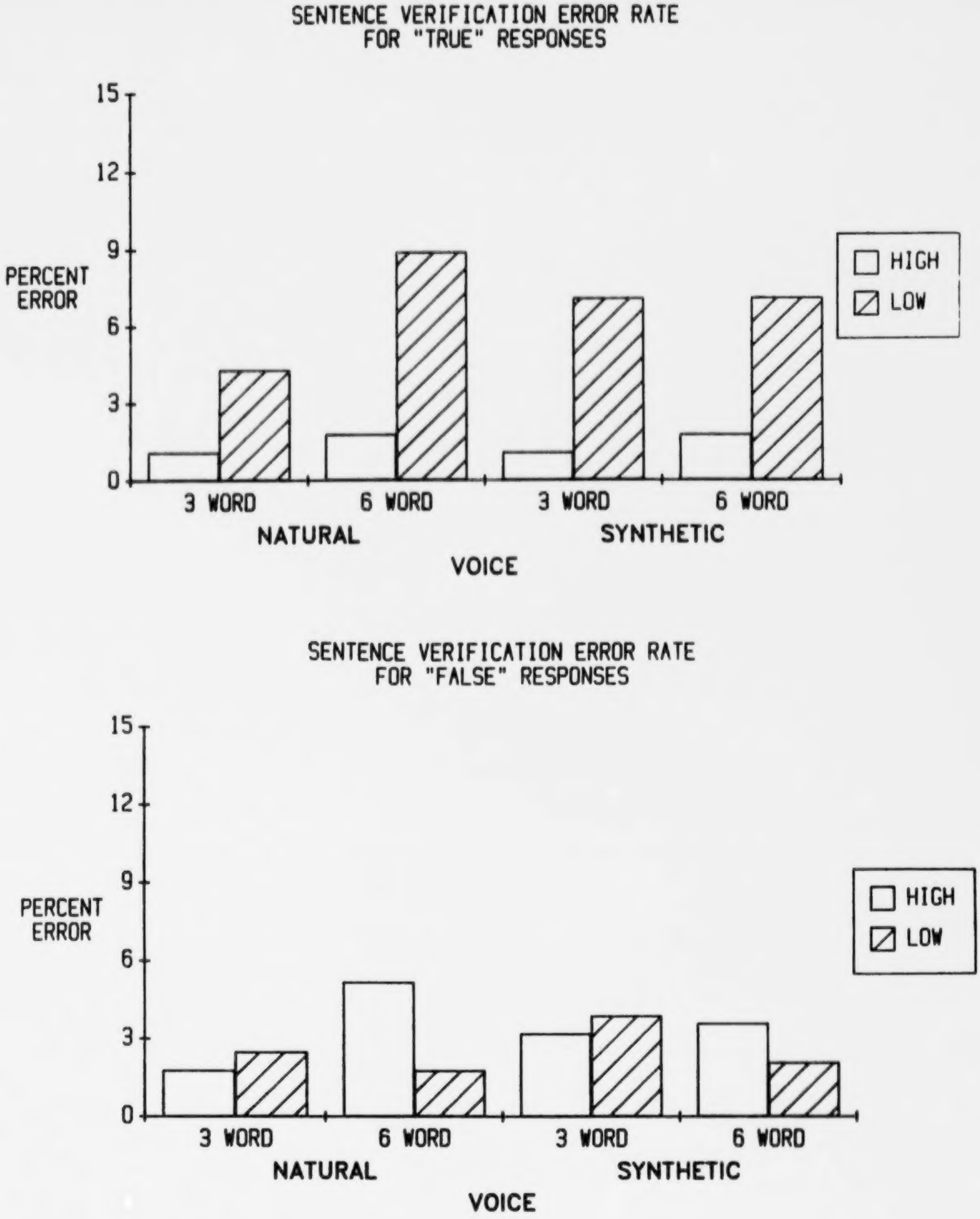
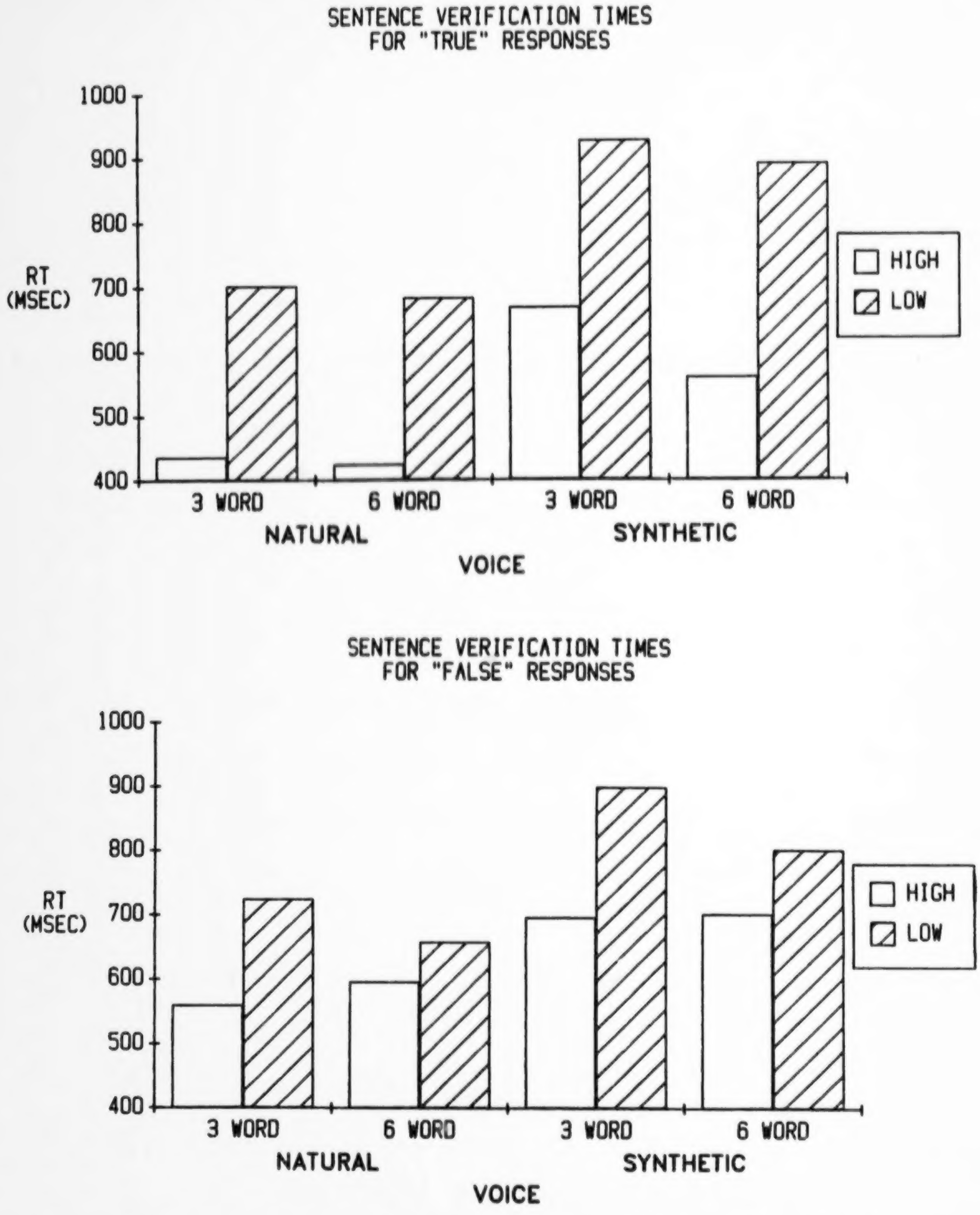


Figure 1. Sentence verification error rates for "True" responses (top panel) and "False" responses (bottom panel) for natural and synthetic speech at each of two sentence lengths. The high-predictability sentences are displayed with open bars; the low-predictability sentences are displayed with striped bars.



SENTENCE VERIFICATION TIMES FOR "TRUE" RESPONSES

SENTENCE VERIFICATION TIMES FOR "FALSE" RESPONSES

Figure 2. Mean sentence verification latencies (in msec) for "True" responses (top panel) and "False" responses (bottom panel) for natural and synthetic speech at each of two sentence lengths. The high-predictability sentences are displayed with open bars; the low predictability sentences are displayed with striped bars. The latencies shown in this figure are based on only those trials in which subjects responded correctly and also transcribed the sentence correctly.

Inspection of both panels in Figure 2 shows several consistent effects for true and false response latencies. First, there is a very prominent effect of sentence predictability on response latency. This may be observed in both panels of the figure. High predictability sentences were responded to much more rapidly than low predictability sentences and this is present for both true and false sentenses, respectively. Table 3 shows the mean latencies for each of the four cells (collapsing across voice) in both conditions of the experimental design.

Insert Table 3 about here

Separate analyses of variance on the true and false responses revealed highly significant effects for sentence predictability, F(1,54) = 121.64, $p < .001$ and F(1,54) = 32.74, $p < .001$, respectively. No interactions were observed in either analysis for sentence predictability.

Second, a consistent effect of voice can be observed in both panels of Figure 2. Natural speech was responded to more rapidly than synthetic speech. Table 4 provides the mean latencies for the four cells (collapsing across predictability) in the experimental design for the true and false responses.

Insert Table 4 about here

Although the figure displays this effect for both true and false responses, separate analyses of variance established that the effect of voice was only significant for the true responses (F(1,54) = 9.07, $p < .004$). The ANOVA for the false responses produced a result that was in the expected direction from the data trends shown in Figure 2, but it did not reach statistical significance (F(1,54) = 3.94, $p < .053$).

None of the other main effects or interactions reached statistical significance in either analysis of true or false responses and no interactions were observed with either of the two main variables (i.e., predictability and voice) that did reach significance. Thus, consistent and reliable differences in verification latencies between natural and synthetic speech were observed even when the sentences were controlled for intelligibility. These results provide evidence against the claim that the observed differences were due to differences in segmental intelligibility between natural and synthetic speech or differences in encoding strategies at the time of input. The 200 msec overall mean difference in the response latencies between natural and synthetic speech found for the true responses suggests that some aspect of the comprehension process other than perceptual encoding is affected by the quality of the acoustic-phonetic input in the speech signal. It is clear from these findings that highly intelligible synthetic speech produced by rule still produces a decrement in performance even when elaborate steps have been taken to experimentally control for differences in the initial level of intelligibility of the stimulus materials. The nature of these differences in the comprehension process will be considered below.

Table 3. Mean Response Latencies for Predictability and Length

TRUE

	HIGH PREDICTABILITY	LOW PREDICTABILITY
3-word	553	815
6-word	492	788

FALSE

	HIGH PREDICTABILITY	LOW PREDICTABILITY
3-word	629	813
6-word	650	730

Table 4. Mean Response Latencies for Voice and Length

	TRUE	
	NATURAL	SYNTHETIC
3-word	570	799
6-word	555	725

	FALSE	
	NATURAL	SYNTHETIC
3-word	643	799
6-word	627	753

General Discussion

The present investigation was designed to study the comprehension process using much more sensitive response measures than have been employed in earlier studies dealing with the perception and comprehension of synthetic speech produced automatically by rule. Using short meaningful three- and six- word sentences that were controlled for segmental intelligibility, we found that response latencies in a sentence verification task were reliably faster for sentences produced using natural speech than the same sentences produced using high-quality synthetic speech generated by DECtalk. Thus, to a first approximation, we were reasonably successful in finding a comprehension task that would reveal meaningful differences in performance between natural speech and very high quality synthetic speech. In the sections below, we offer an account of these findings in terms of earlier work using the sentence verification task to study language comprehension processes.

As we noted in the introduction to this report, previous studies on the comprehension of synthetic speech have consistently failed to find reliable differences in performance between natural speech and several kinds of low-quality synthetic speech. Such findings have appeared anomalous to us because other measures of phoneme perception, word recognition and sentence transcription all reliably discriminated not only between natural and synthetic speech but more importantly, between different kinds of synthetic speech ranging from high quality systems such as DECtalk to very poor quality systems such as ECHO (see Greene, Logan and Pisoni, 1986). We raised a number of criticisms about the specific experimental procedures used in these earlier studies, including some of own research, and we offered several suggestions as fruitful alternatives to pursue in future work on this problem. The present experiment which used a sentence verification task to study comprehension was specifically designed with these criticisms in mind.

In addition to finding differences in the verification latencies between natural and synthetic speech, we also observed a reliable effect of sentence predictability on response latencies. This effect was found for both true and false responses and was extremely robust. High predictability sentences were consistently responded to more rapidly than low predictability sentences. To our surprise, however, we failed to find any reliable effects of sentence length on verification latencies. We failed to find any interactions among the three experimental variables manipulated in this experiment. Contrary to our original expectations, we did not observe the predicted interaction between voice and sentence predictability which would have demonstrated differential effects of predictability on the synthetically produced sentences. Precisely why we failed to find this result is unclear at this time. Several suggestions will be considered below. Additional experimental manipulations will be needed to determine the locus of the observed voice effect in the language processing system. For the present time, however, the absence of the predicted interaction between voice and sentence predictability is an important finding that merits further attention.

The present results differ from our earlier study (see Manous et al., 1985) in a number of respects that are important to consider at this point. In the original SVT study, we found that sentence verification error rates and response latencies were strongly related to the segmental intelligibility of the particular text-to-speech system under study. However, differences in segmental intelligibility among the text-to-speech systems varied quite widely and therefore the observed differences in the verification test could be attributed to a variety of factors among which might be real differences in the comprehension process itself or simply differences in the initial levels of intelligibility of the systems. As it stands, our earlier study could not discriminate between the source or sources of the observed differences in either the verification error rates or the response latencies. Differences in the verification error rates suggest, however, that subjects probably did have difficulty encoding some of the sentences, particularly those produced by the low-quality systems. Our analyses of the transcription data, collected after each sentence was verified, further suggested that this was a reasonable account of the differences. Thus, the most parsimonious explanation of the results of our earlier study was that the observed differences were probably due to difficulties at the time of encoding because the initial level of intelligibility of the systems varied so widely (see Greene, Logan and Pisoni, 1986, for a summary of the intelligibility data for these systems).

With regard to the outcome of the present study, such an explanation would be difficult to find support for because the segmental intelligibility of the test sentences was very carefully controlled before the experimental data were collected. Moreover, the observed error rates for the verification responses were extremely low and no reliable differences could be observed in the pattern of the errors across the experimental conditions. In short, subjects did not have difficulty perceiving the sentences. They did have difficulty, however, in determining whether the sentences were true or false. This decision required subjects to understand the meaning of the sentences and to respond appropriately.

The results of the present study suggest that several additional factors related to processing operations involved in language comprehension may be responsible for the observed differences in verification latencies. Because the test sentences were matched on segmental intelligibility, the present findings suggest that in addition to differences in segmental intelligibility, differences also exist in comprehension between natural speech and high quality synthetic speech generated by DECtalk and that these differences are above and beyond differences related to the intelligibility of speech as measured by traditional types of transcription tests or MRT scores. Whatever the precise locus of the differences, the present findings demonstrate that segmental intelligibility is not sufficient to account for the pattern of response latencies observed in the present study.

In this connection, it is useful to consider briefly the findings of a recent study carried out by Pisoni and Dedina (1986) who used a sentence verification task with natural speech that had been processed using three quite different digital encoding algorithms. Despite the fact that standard tests of segmental intelligibility using the MRT revealed only very small differences in performance among the three vocoders, the results of the verification task revealed quite robust and consistent findings which could be related directly to the data rate of the processing algorithms. Latencies were fastest and error rates were lowest for the 16kbps CVSD algorithm, followed by the 9.6kbps TDHS/SBC algorithm, and finally the 2.4kbps LPC-10 algorithm. Thus, it is reasonable to conclude from the present findings using synthetic speech and the recent data of Pisoni and Dedina using vocoded natural speech, that traditional intelligibility tests may simply be insensitive to important differences that are, in fact, present in the speech waveform and apparently affect the listener's performance in understanding the content of the message and responding appropriately to the truth value of the utterance.

To determine the locus of the observed differences, it is necessary to examine the comprehension process in somewhat greater detail within the framework of a specific model. In recent studies of language comprehension, specifically, in experiments on sentence verification, it has become common to view comprehension as a "process" and to divide that process into a series of processing stages. Clark and Clark (1977) describe a generic sentence verification model of sentence comprehension with the following four stages: Stage 1 represents the interpretation of the sentence; Stage 2 represents the relevant external or internal evidence; Stage 3 compares the representations from Stages 1 and 2; and Stage 4 responds with the answer computed at Stage 3. According to this model, each stage has one or more cognitive operations and each operation takes some amount of processing time to complete. In applying this model to the sentence verification task used in the present investigation, it is assumed that listeners begin at Stage 1 and by the time they get to Stage 4 they are able to respond either "true" or "false."

At Stage 1, listeners construct some internal representation of the meaning of the sentence. For present purposes, the exact nature of the internal representation is not important. What is important, however, is that this stage of the model involves the encoding of the input sentence into a format that can be used in Stage 3, the comparison stage. At Stage 2, listeners represent the relevant external or internal evidence in the same format as that used in Stage 1. In the present investigation, this information is retrieved from the listener's knowledge stored in long-term memory. This knowledge is assumed to be available because subjects can easily verify simple statements concerning facts and knowledge that they know from their past experiences. Thus, subjects have to retrieve relevant information from long-term memory and represent that information in some format that can also be used in Stage 3 of the model. At Stage 3, the comparison stage, listeners compute a "truth index," which will be used to select the correct response. According to Clark and Clark (1977), the comparison process at Stage 3 consists of two rules. First, listeners start with the truth index set to true. Second, they compare the two representations to determine if they match. If the two representations match, the truth index is left alone. If the two representations do not match, the truth index is changed to the opposite setting, false. Thus, the comparison stage of this model is central to determining the truth value of the two representations. Finally, at Stage 4, the listener examines the final truth index and responds accordingly. If the truth index is true the subject responds "true," and if the truth index is false, the subject responds "false." In addition to assuming that each of these operations takes up some processing time, the model also assumes that the comparison process relies on the congruence of the two representations computed at Stages 1 and 2. Previous studies have shown that it is easier to decide that two representations match than to decide if they mismatch.

Considering the framework of the verification model outlined above, it is possible to speculate about the locus of the differences observed in the present study. Although we have tried to argue that the differences found in the present study are not due to factors related to segmental intelligibility, and we have been cautious not to over-interpret the results of the present study, it is still possible that our findings are due to some aspect of the perceptual encoding process either at the time of input or at the time the initial representation of the meaning of the test sentence is constructed at Stage 1 of the model. Thus, the initial representation of the synthetic speech at Stage 1 in the model outlined above may be degraded or noisy in some way relative to natural speech. Because standardized tests of speech intelligibility are not performance limited, that is, subjects are not typically required to respond rapidly in these tasks, it is quite possible that transcription scores and MRT results typically obtained with high-quality synthetic speech or digitally encoded natural speech are much too insensitive to pick up any of the differences that are localized at Stage 1 of the verification model, the stage at which the initial representation of the sentence is constructed from the speech waveform.

If this line of reasoning is correct, or nearly so, it would imply that the initial representation of the test sentence is encoded in a format that contains some information about the acoustic-phonetic quality or attributes of the input signal. Put another way, some property or set of properties related to the perceptual analysis of the speech waveform and/or its segmental representation are passed along or "propagated" up the processing system to higher and progressively more abstract levels of the comprehension process. One consequence of this account of our findings would be a general slowing up of all processing activities in the comprehension task under these conditions. This result would not be affected by other experimental manipulations such as sentence predictability or sentence length that may have their effects localized at Stages 2 or 3 in the verification model. Indeed our failure to find an interaction between voice and predictability would be consistent with this explanation and would imply that the locus of the predictability manipulation occurs somewhere later in the comprehension process than the voice manipulation, perhaps at Stage 2 where relevant information is retrieved from long-term memory or possibly at Stage 3, where the two representations are compared.

Similar findings have been reported by Pisoni (1981) and by Slowiaczek and Pisoni (1982) who used lexical decision and naming tasks to study the perception of isolated words that were either natural speech or synthetic speech generated by the MITalk text-to-speech system. Both studies found longer response latencies for synthetic speech compared to natural speech. However, no interactions were observed with any of the other experimental variables suggesting, as we found in the present study, that the locus of the effects of the voice manipulation appear to be at either the initial stage of perceptual encoding or the development of some initial representation of the input signal that will be used in the comparison process in verification. Without further studies utilizing additional experimental manipulations, it is not possible to decide on which of these alternatives is the correct account of the present results. However, it is clear that we have found robust effects of the voice manipulation on some selected aspects of the comprehension process that appear to be separable from effects related to segmental intelligibility. The subjects in the present experiment had no difficulty whatsoever in perceiving the words or sentences or responding to the truth value of the meaning of the sentences. Our primary finding was that the response latencies were considerably shorter when the sentences were natural speech than when the sentences were produced with high-quality synthetic speech generated by DECtalk.

In summary, the results of the present investigation demonstrate that some aspect of the comprehension process, either the encoding of the initial representation or the comparison process, is affected by the quality of the acoustic-phonetic input in the speech signal. Using short meaningful sentences that were controlled for segmental intelligibility, we found that verification latencies were reliably shorter for natural speech than high-quality synthetic speech produced by rule using DECTalk. Further studies are currently underway to identify the locus of these effects in the human information processing system and to specify the nature of the processing operations that are affected by these differences in the initial sensory-based input in the speech signal. The results of the present study taken together with the earlier findings of Manous et al., (1985) and the more recent data of Pisoni and Dedina, (1986) demonstrate that the SVT appears to be a useful and extremely sensitive tool for investigating differences in the comprehension process when initial differences in intelligibility are small or nonexistent. The results also demonstrate the need to develop new and more sophisticated measures of sentence comprehension that can be used to study speech communication at processing levels above and beyond those typically indexed through transcription tasks or traditional forced-choice intelligibility tests such as the Modified Rhyme Test or Diagnostic Rhyme Test.

References

Allen, J. (1981). Linguistic-based algorithms offer practical text-to-speech systems, *Speech Technology*, 1(1), 12-16.

Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, *99*, 199-220.

Clark, H. H., & Clark, E. V. (1977). *Psychology and Language*. New York: Harcourt Brace.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240-247.

Collins, A. M., & Quillian, M. R. (1970). Facilitating retrieval from semantic memory: The effect of repeating part of an inference. In A. F. Sanders(Ed.), *Attention & Performance III*, Amsterdam: North Holland Publishing Co.

Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology*, *92*, 149-154.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, *58*, 955-991.

Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, *4*, 107-111.

Gough, P. B. (1966). The verification of sentences: The effects of delay of evidence and sentence length. *Journal of Verbal Learning and Verbal Behavior*, *5*, 492-496.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers*, *17*, 100-107.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, *37*, 158-166.

Jenkins, J. J. & Franklin, L. D. (1981). Recall of passages of synthetic speech. Paper presented at the Psychonomics Society Meeting, November, 1981.

Larkey, L. S., & Danly, M. (1983). Fundamental frequency and sentence comprehension. *MIT Speech Group Working Papers, Vol. II*.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors*, *25*, 17-32.

Manous, L.M., Pisoni, D. B., Dedina, M. J., & Nusbaum, H. C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. Research on Speech Perception Progress Report No. 11. Bloomington, IN: Indiana University.

McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Naval Research Lab (9-9-76).

Moody, T. S. & Joost, M. G. (1986). Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. Proceedings of the Voice Input/Output Society, Alexandria, VA, 1986.

Nixon, C. W., Anderson, T. R., & Moore, T. J. (1985). The perception of synthetic speech in noise. In R. Salvi, D. Henderson, R. P. Hamernik, & V. Coletti (Eds.), Applied and Basic Aspects of Noise Induced Hearing Loss. NY: Plenum, in press.

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. Behavior Research Methods, Instruments, & Computers, 17, 235-242.

Nye, P. W., & Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural control (Modified Rhyme Test results). Haskins Laboratories Status Report on Speech Research, SR-33, 77-91.

Pisoni, D. B. (1979). Some measures of intelligibility and comprehension. In J. Allen (Ed.), Conversion of Unrestricted English Text to Speech. Cambridge: Cambridge University Press (forthcoming).

Pisoni, D. B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America, 70, S98.

Pisoni, D. B. (1982). Perception of speech: The human listener as a cognitive interface. Speech Technology, 1, 10-23.

Pisoni, D. B., & Dedina, M. J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report. Research on Speech Perception Progress Report No. 12. Bloomington, IN: Indiana University.

Pisoni, D. B., & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In 1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing (pp. 572-575). New York: IEEE Press.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. Proceedings of the IEEE, 11, 1665-1676.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. Human Factors, 27, 395-408.

Slowiaczek, L. M., & Pisoni, D. B. (1982). Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America, 71, S95-96.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Perceptual Learning of Synthetic Speech Produced by Rule*

Steven L. Greenspan, Howard C. Nusbaum, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This research was supported in part by Air Force Contract No. AF-F-33615-83-K-0501 with the Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH; in part by NIH Grant NS-12719; and in part by NIH Training Grant NS-07134 to Indiana University in Bloomington. The authors thank Kimberly Baals and Michael Stokes for their valuable assistance in collecting and scoring the data; and Robert Bernacki, Michael Dedina, and Jerry Forshee for technical assistance.

Abstract

To examine the effects of stimulus structure and variability on perceptual learning, we compared transcription accuracy before and after training for synthetic speech produced by rule. In the first experiment, subjects were trained either with isolated words or fluent sentences of synthetic speech. In addition, subjects were presented either with novel stimuli during each training session or with a fixed list of stimuli that was repeated. A control group received no training. The results indicate that training with isolated words only increased the intelligibility of isolated words; however, training with sentences increased the intelligibilty of both isolated words and sentences. This finding suggests that listeners do not segment words in fluent speech by recognizing one word at a time. Furthermore, subjects who were trained on the same stimuli every day improved as much as the subjects who were given novel stimuli on each day of training. This finding was further investigated in a second experiment in which the size of the repeated stimulus set was reduced to enable subjects to quickly and completely learn the items in the training set. Under these conditions, subjects trained with repeated stimuli did not generalize to novel stimuli on the post-training test as well as subjects trained with novel stimuli. Taken together, the results on the effects of training with repeated items versus novel items suggests that perceptual learning depends upon the degree to which the training stimuli characterize the underlying structure of the full stimulus set. Variability of the stimulus ensemble aids generalization to novel tokens.

Perceptual Learning of Synthetic Speech Produced by Rule

In general, there are two basic experimental paradigms within cognitive psychology for investigating learning. The first, usually identified with Ebbinghaus (1964/1885), is concerned with learning a fixed set of stimuli through repetition. This approach emphasizes the acquisition, retention, retrieval, and representation of specific facts (see Kolers & Roediger, 1984) and therefore may be said to investigate rote or pattern-specific learning. Recent evidence suggests that even a single prior occurrence of a degraded stimulus during the study phase of an experiment can increase the probability of its perceptual identification in a subsequent test (Jacoby, 1983). This result was obtained even when only 10% of the test items had been seen previously, and when the two presentations were separated by several days and by other lists of stimuli. Moreover, the advantage of a repeated test item over new test items was evident despite nonspecific practice effects for identifying both new and repeated items. This suggests that subjects encode specific episodic traces of stimuli that can be used to aid subsequent perceptual analysis (Jacoby, 1983).

The second paradigm, often associated with James (1890), Woodworth (1938), Bartlett (1932), and Posner and Keele (1968), has been concerned more with the development of abstract characterizations of stimulus information. Typically, in this approach, subjects are first presented with a series of stimuli in which no stimulus is repeated. During this phase of the experiment, the subject is expected (or explicitly trained) to abstract some characteristics from these stimuli. Afterwards, generalization to novel stimuli is examined. Thus, this second paradigm focuses on the ability to classify or respond to novel instances of a category. Under this experimental procedure, learning is usually characterized by the development of abstract concepts, schemata, or rules that represent the underlying structure of each stimulus or an entire stimulus set. The difference between these two paradigms highlights important issues underlying any account of perceptual learning: What are the effects of stimulus repetition, stimulus variability, and stimulus structure on perceptual learning?

Current evidence suggests that presenting subjects with stimuli that vary around a prototype encourages the development of a mental representation of the prototype. For example, Posner and Keele (1968) presented subjects with random distortions of four dot patterns that served as the prototypes of four categories. Although subjects never saw the prototypes during the study phase of the experiment, the results suggested that the prototypes and the previously-seen training patterns were equally easy to classify, and were both more quickly and more accurately classified than novel category members. In addition, the results reported by Posner and Keele (1968) indicated that as the variability of the training set increased, so did the abililty to classify new, highly distorted variations of the prototype. Thus, subjects in the Posner and Keele experiment appeared to be abstracting information about category prototypes and the variability of the category space, as well as retaining information about the exemplars that were presented during training.

In a related study that required subjects to learn names for pictures of individuals (Dukes & Bevan, 1967), subjects received training with either one exemplar per category or with several exemplars per category and the number of training trials was kept constant across conditions. Thus, subjects saw either many repetitions of the same exemplar or few repetitions of several exemplars. During the test phase, subjects were presented with old and new exemplars. The results suggested that training by repeated exposure to a single exemplar enhanced the ability to classify that exemplar in a subsequent test relative to training with several different exemplars. However, training with several examplars increased the ability to classify novel category exemplars. Moreover, consistent with the results of Posner and Keele (1968), the exemplars that were previously seen were more accurately classified than new exemplars for both groups of subjects. (This type of result was also obtained by Kolers, 1976.) These results suggest that as the number of training exemplars for a particular category increases, the ability to classify new instances of that category will also increase.

However, this conclusion must be qualified in light of results obtained in several artificial-language learning studies. In this research, subjects are presented with a set of patterns that were generated by several phrase structure and lexical insertion rules. Subjects are told that they are being presented with a subset of an artificial language. After exposure to the training set, subjects are presented with novel instances and must either produce grammatical sentences or make grammatical judgements. Studies by Nagata (1976) and Palermo and Parish (1971) suggest that if the total number of training presentations is kept constant, presenting few exemplars many times or many exemplars few times will have equivalent effects, as long as the training stimuli sufficiently characterize the set of possible grammatical sentences.

Thus, in these artificial-language learning studies, the arbitrary variability of the training set, per se, did not affect transfer to novel test stimuli. Moreover, these results were obtained with very different grammatical systems and whether or not a semantic-referential framework had been provided. The difference between the conclusions suggested by these langauge learning studies and those of Dukes and Bevan (1967) and Posner and Keele (1968) may be due to the difference in the stimulus materials. In the Posner and Keele study, different exemplars of a prototype were stochastically related to the prototype. In contrast, the grammatical sequences presented in the artificial-language learning studies were well-defined realizations of a set of coherent grammatical rules. Thus, the importance of variability may be minimized when the exemplar space is highly structured and can be well-defined by a set of abstract rules. A few well-chosen exemplars may be sufficient to induce the abstraction of the complete structure of the rules to produce generalization learning.

Considered in this context, speech signals provide an especially interesting and important class of stimuli for studying the effect of stimulus variability on perceptual learning. The acoustic-phonetic variability of the speech signal has been well-established (e.g., Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). To a large extent, this variability has often been viewed as noise that must be "stripped away" from the speech signal in order to reveal invariant phonetic structures (e.g., Stevens & Blumstein, 1978). However, acoustic-phonetic variability may also be viewed as a highly structured and coherent source of information about the talker, phonetic context, and speaking rate that is treated by the listener as information rather than noise (Elman & McClelland, 1986). This structural coherence arises, in part, because the sources of variability in speech production are not arbitrary. Factors such as speaking rate, vocal tract size and shape, and phonetic context have regular, well-defined and physically specifiable effects upon the acoustic realization of a phonetic segment. Thus, although the production of a phoneme is strongly determined by the surrounding phonetic context through coarticulation, these coarticulatory effects are, in principle, computable, and therefore may serve as information about the segment and the context in which it is produced (cf. Cole, Rudnicky, Zue, & Reddy, 1980; Greene, Pisoni, & Carrell, 1984).

Although it is difficult, if not impossible, to segment the speech signal into linguistically defined units that are independent of one another, and although there is no simple one-to-one correspondence between the acoustic structure of the speech signal and the perceived phonetic structure, listeners nonetheless show great facility for quickly adapting to novel acoustic-phonetic variation. For example, listeners appear to have little difficulty understanding novel speakers of their langauge despite substantial variations in vocal tract size and shape, and manner of speaking. Moreover, although the speech of some non-native speakers may be difficult to comprehend because of violations in the phonetic, phonological, syntactic, and prosodic regularities, word recognition and sentence comprehension generally improves as the listener gains experience with the non-native speech. Thus, on the basis of the previously described artificial-language learning studies, if the variability of speech is systematic, the ability to adapt to distorted speech may depend not upon the relative number of novel exemplars experienced during learning, but rather upon the degree to which the training set is sufficiently representative of the underlying structure of the distorted speech. This is the issue that we investigated in the present experiments.

In order to investigate the effects of stimulus structure, repetition, and variability on perceptual learning of speech, we used a rule-based text-to-speech system to produce synthetic speech of relatively low intelligibility. Unlike natural speech, synthetic speech produced by rule has an impoverished acoustic-phonetic cue structure that incorporates only those acoustic cues that can be easily described by a small set of phonetic implementation rules (cf. Liberman, Ingeman, Lisker, Delattre, & Cooper, 1959). In contrast, natural speech provides a rich set of redundant cues for each distinctive phonemic contrast in the language (see Lisker, 1978). There is now considerable evidence that this difference in acoustic-phonetic structure between synthetic and natural speech has important perceptual consequences for the human listener (Nusbaum, Dedina, & Pisoni, 1984; Yuchtman, Nusbaum, & Pisoni, 1985).

To take one example, the pattern of perceptual confusions observed for synthetic syllables is quite different from the pattern of confusions observed for natural speech degraded by noise, even when the intelligibility of the natural speech in noise is comparable to the intellibility of the synthetic speech (Nusbaum, et al, 1984; Yuchtman et al. 1985). Nusbaum et al. (1984) and Yuchtman et al. (1985) suggest that the difficulties incurred in perception of synthetic speech are due primarily to the use of minimal acoustic cues to synthesize phonetic contrasts in current text-to-speech systems. Indeed, in some contexts, these cues are insufficient to distinguish phonetic segments in particular contexts; in other contexts, the cues may actually be inappropriate for the intended phonetic segment. Thus, improvements in recognizing synthetic speech produced by a text-to-speech system may require learning new perceptual mappings between acoustic cues and phonetic categories. Furthermore, it may also require the listener to learn to attend to and discriminate acoustic information that is not normally used to distinguish phonemes in the listener's native language.

Whether listeners can, in fact, acquire new mappings between acoustic information and phonetic labels is a matter of some controversy (Pisoni, Aslin, Perey, & Hennesy, 1982). For example, Strange and Jenkins (1978) reviewed a number of studies that attempted to train subjects to identify and discriminate nonphonemic differences in voice-onset time. On the basis of these studies, Strange and Jenkins argued that the use of laboratory training techniques with adult subjects is generally ineffective for improving the discriminability of phonetic contrasts that are not phonemically distinctive in the subject's native language. However, Pisoni et al. (1982) provided evidence that the ability to discriminate nonphonemic differences in voice-onset time can be acquired after a short training period, if appropriate training procedures are used (e.g., additional response categories, immediate feedback, and a brief exposure to the stimulus-response set prior to training). Thus, listeners are able to learn new relationships between acoustic information and a set of phonetic labels.

More recently, Schwab, Nusbaum, and Pisoni (1986) have demonstrated that moderate amounts of training with low-intelligibility synthetic speech will improve word recognition performance for novel stimuli generated by the same text-to-speech system. Schwab et al. trained subjects by presenting synthetic speech followed by immediate feedback in recognition tasks for words in isolation, in fluent meaningful sentences, and in fluent semantically anomalous sentences. Subjects trained under these conditions improved significantly in recognition performance for synthetic words in isolation or in sentence contexts compared to subjects that either received no training or received training on the same experimental tasks with natural speech. Thus, the improvement found for subjects trained with synthetic speech could not be ascribed to mere practice with or exposure to the test procedures. In addition, a follow-up study indicated that the effects of training with synthetic speech persisted even after six months. Thus, training with synthetic speech produced reliable and long-lasting improvements in perception of words in isolation and words in fluent sentences.

The finding that recognition improved both for words in isolation and for words in fluent speech is of some theoretical importance and interest because recognizing words in fluent speech presents a problem that is not present when words are presented in isolation: The context-conditioned variability between words and the lack of independence between adjacent acoustic segments leads to enormous problems for the segmentation of speech into psychologically meaningful units that can be used for recogniton. In fluent, continuous speech it is extremely difficult to determine where one word ends and another begins using only acoustic criteria (Pisoni, 1985; although cf. Nakatani & Dukes, 1977).

McClelland and Elman (1986; see also Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; Reddy, 1976) have recently proposed a model of speech perception called Trace, in which word segmentation is a direct consequence of word recognition. In this model, there is a lexical basis for segmentation such that recognition of the first word in an utterance determines the end of that word as well as the beginning of the next word in the utterance. Consistent with this proposal, there are no explicit mechanisms in Trace for segmenting words prior to recognition. Although Trace was not intended to address the issues surrounding perceptual learning, their model suggests that training subjects with isolated words generated by a synthetic speech system should improve the recognition of words in fluent synthetic speech and, conversely, training with fluent synthetic speech should improve performance on isolated words. According to Trace, if listeners recognize isolated words more accurately, word recognition in fluent speech should also improve since, in this model, perception of words in fluent speech is a direct consequence of the same recognition processes that operate on isolated words.

However, recent evidence from studies using visual stimuli suggests that differences in the perceived structure of training stimuli may lead to the acquisition of different types of perceptual skills. Kolers and Magee (1978) presented inverted printed text and in a training task instructed subjects either to name the individual letters in the text or to read the words. After extensive training, subjects were found to have improved only on the task for which they received training: Attending to letters improved performance with letters, but had little affect on reading words; similarly, attending to words improved performance with words but had little affect on naming letters. However, results for visual stimuli may not necessarily apply to speech because of the substantial differences that exist between spatially distributed, discrete printed text and temporally distributed, context-conditioned speech.

To summarize, there are several basic issues that are directly relevant to understanding perceptual learning, in general, and to understanding the perceptual learning of speech, in particular. These issues concern the effects of stimulus repetition, stimulus variability, and stimulus structure on perceptual learning of speech. Previous research has demonstrated that repeating presentations of a stimulus have powerful effects on subsequent perception of that stimulus (e.g., Jacoby, 1983). However, it is important to understand the relationship between these repetition effects and the type of generalization learning that is so important for acquiring new perceptual skills. According to Shiffrin and Schneider (1977), the automatization of a perceptual process is the direct consequence of simple repetitions of fixed stimulus-response mappings. However, this view of learning may not be sufficient to generalize beyond highly restricted tasks using simple stimuli in very impoverished environments to complex perceptual-motor skills that must not only be well-learned, but must also deal with environmental variability. Except under laboratory conditions, it is very seldom that listeners hear precisely the same utterance with the same acoustic structure more than once. However, listeners do learn to accomodate the prodigious variability in the structure of speech in a highly skilled manner. The present studies were directed at investigating perceptual learning of synthetic speech generated by rule. In this research, we were particularly interested in comparing the effects on perceptual learning of repeating a fixed set of training stimuli with the effects of continuously presenting novel training stimuli.

A second goal of the present research was to investigate how the pattern structure of stimulus materials encountered during learning affects generalization to other novel stimuli. Theories of speech perception that posit word segmentation as a consequence of word recognition predict that any improvement in word recognition should also improve recognition of words in fluent speech. However, research on perceptual learning of inverted printed letters and words suggests that there should be little generalization from perceptual learning on one type of pattern to recognition of another type of pattern (e.g., Kolers & Magee, 1978). We trained subjects with isolated words or sentences produced by a text-to-speech system to investigate generalization of learning across different stimulus materials and pattern structures.

Experiment 1

The first experiment was carried out to examine the effects of training based on different stimulus materials. We trained subjects with either isolated words or with fluent sentences (but not both), and then examined whether each type of training would improve recognition of novel words presented in isolation and in fluent sentences. If word segmentation is a direct consequence of word recognition (as suggested by Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986), then improvements in isolated word recognition should produce better word recognition in fluent speech. Alternatively, different linguistic materials may require the acquisition of different perceptual skills (as occurred in Kolers & Magee, 1978), so that training with isolated words might only improve performance on isolated words and training with fluent sentences might only improve performance on fluent sentences. Finally, a third possibility is that some complex linguistic materials may subsume other, structurally simpler linguistic materials. On the one hand, sentences necessarily contain words so training with fluent sentences may improve performance with isolated words as well as with words in fluent speech. On the other hand, training with isolated words might not generalize to recognition of words in fluent sentences because recognition of words in sentences may require skills that cannot be acquired from experience with isolated words alone.

We should note here that the stimulus materials used in Experiment 1 were produced by the Votrax Type-'N-Talk text-to-speech system. The fluent synthetic speech produced by the Votrax system has certain characteristics that are useful for testing the hypotheses under question. An LPC analysis (see Markel & Gray, 1976) of the Votrax-produced stimuli indicated that a word excised from a fluent sentence produced by Votrax was identical to the same word produced in isolation. Both words have identical formant structures and equivalent pitch and amplitude contours. Thus, the fluent sentences produced by the Votrax system are merely end-to-end concatenations of individual words (with no pauses or coarticulation phenomena between words). The Votrax system does not introduce any systematic acoustic information in its fluent speech that is not already present in its productions of isolated words so there are no observable sentence-level effects on phonetic segements. Therefore, the Votrax speech provides an excellent set of stimuli for testing the claim that word segmentation is a direct consequence of word recognition. Since a sentence produced by the Votrax system is equivalent to a sequence of isolated words, improvements in recognizing isolated words should directly generalize to recognizing words in sentences.

In addition to examining the influence of stimulus structure on perceptual learning in this experiment, the effect of stimulus repetition was also investigated. Some subjects received novel stimuli throughout training, so that they never heard any stimulus more than once. Other subjects received a fixed set of stimuli that was repeated several times during training. Thus, some subjects were always trained on new words or sentences, while other subjects heard the same words or sentences over and over again. Both groups were tested on novel stimuli before and after training to examine generalization learning to novel words and sentences. Based on the Posner and Keele (1968) and Dukes and Bevan (1967) experiments, subjects trained on novel exemplars should show more improvement than those trained with repeated exemplars, because the novel training set provides a more variable sample of speech than is provided by the repeated training set. However, different predictions follow from the artificial-language learning studies of Nagata (1976) and Palermo and Parish (1971). On the basis of these studies, we can predict that since the variability of speech is lawful and if the repeated training set sufficiently characterizes the underlying rule structure of the speech, there should be no performance difference between subjects receiving a repeated training set and those presented with a novel training set (as long as both sets of subjects receive an equal number of exemplars).

Method

Subjects. Sixty-six naive subjects participated in this experiment. All were students at Indiana University and were paid four dollars for each day of the experiment. All subjects were native speakers of English, and reported no previous exposure to synthetic speech and no history of a hearing or speech disorder. All subjects were right-handed and were recruited from a paid subject pool maintained by the Speech Research Laboratory of Indiana University.

Materials. All stimulus materials were produced by a Votrax Type-'N-Talk text-to-speech system controlled by a PDP-11 computer. The Votrax system was chosen for generating words and sentences because of the relatively poor quality of its segmental (i.e., consonant and vowel) synthesis. Thus, the likelihood of ceiling effects in word recognition were minimized. The synthetic speech stimuli used in the present study were a subset of the stimuli developed and used by Schwab et al. (1985) to insure comparability between experiments.

The stimuli were produced by the Votrax Type-'N-Talk system controlled by a PDP-11/34 computer. All stimulus materials were initially recorded on audio tape. After the audio recordings were made, the stimulus materials were sampled at 10 kHz, low-pass filtered at 4.8 kHz, digitized through a 12-bit A/D converter, and stored in digital form on disk with the PDP-11/34 computer. The stimuli were presented in real time at 10 kHz through a 12 bit D/A converter and low-pass filtered at 4.8 kHz. Four sets of stimulus materials were used in this experiment:

1. PB Lists. The first set of stimuli consisted of 12 lists of 50 monosyllabic, phonetically balanced (PB) words. These lists were a subset of the 20 lists originally designed for testing speech intelligibility (Egan, 1948). These stimuli were used during testing and training procedures. These lists are considered to be phonetically balanced because their phonetic composition provides a reasonable approximation of the relative frequencies of phonemes occuring in English (Egan, 1948).

2. MRT Lists. The second set of materials consisted of four lists of 50 monosyllabic consonant-vowel-consonant words taken from the Modified Rhyme Test (MRT) developed by House, Williams, Hecker, and Kryter (1965). Although both the PB lists and the MRT lists contain monosyllabic words, these lists were used in different tasks. The PB lists were used in an open-response set transcription task, while the MRT lists were used in a closed-response, six-alternative forced-choice procedure. Moreover, in combination with the closed-response set procedure, the MRT lists were designed to examine segmental phonetic intelligibility (as opposed to lexical intelligibility). On each trial, the six alternatives differ by a single consonant in one location. The alternatives either had identical beginnings (i.e., the same initial consonant and vowel) or identical endings (i.e., the same vowel and final consonant). As a consequence, the MRT examines the ability to discriminate consonants. The MRT lists were only presented during the pre-training and post-training test sessions to observe the effects of different training conditions on perceptual learning.

3. Harvard Sentences. The third set of stimuli consisted of 10 lists of 10 Harvard psychoacoustic sentences (Egan, 1948; IEEE, 1969). These are normal, meaningful, English sentences. Each sentence contains five key words plus a variable number of function words, arranged in a variety of syntactic structures.

4. Haskins Sentences. The fourth set of materials consisted of 10 lists of 10 syntactically normal, but semantically anomalous sentences that had been developed at Haskins Laboratories (Nye & Gaitenby, 1974). Each Haskins sentence contains four high-frequency monosyllabic key words presented in the following syntactic structure: "The (adjective) (noun) (verb, past tense) the (noun)". These sentences minimize word identification based on semantic cues although syntactic cues are present.

Design. The entire experiment was conducted in six one-hour sessions on different days. Five groups of subjects were tested on the first and last day of the experiment. The four intervening days were used to provide training for subjects in four of the five groups. A weekend separated the pre-training test session (on Day 1) from the first day of training (Day 2). All groups were treated similarly during the pre-training and post-training test sessions (Days 1 and 6). However, each group was treated differently during training.

One group of subjects (the novel-word group) received a different set of isolated words on each day of training, while a second group (the novel-sentence group) received a different set of fluent sentences each day of training. In the two novel-stimulus conditions, subjects were continually exposed to new stimuli. A third group (the repeated-word group) received a fixed set of isolated words on every day of training. Similarly, a fourth group of subjects (the repeated-sentence group) received a set of fluent sentences on the first day of training that was repeated for all other training sessions. Thus, in the repeated-stimulus conditions, subjects were presented with novel stimuli only on the two test days (Days 1 and 6), and on the first day of training. (The stimulus set presented on the first day of training was subsequently repeated on the remaining training days.) The last group of subjects (the control group) received no training and provided a baseline against which the performance of the other four groups could be compared.

Procedure. All stimuli were presented to subjects in real time under computer control through matched and calibrated TDH-39 headphones at 77 dB SPL measured using an RMS voltmeter. Before each experimental session, signal amplitudes were calibrated using the same isolated word (from a PB list). Subjects were tested in groups with a maximum of six subjects per group.

Testing Procedure. All subjects were tested during the pre-training and post-training test sessions using the same procedures and order of tasks. Each test session lasted about one hour. In each session, recognition performance for isolated words was tested with an open-response set procedure using PB lists (Egan, 1948) and with a closed-response set procedure using the MRT (House et al., 1965). In the PB task, subjects were presented with two lists of 50 monosyllabic words, one word at a time. After each word was presented, subjects were asked to write the English word that they heard. They were encouraged to guess if they were uncertain about the identity of the word. After writing a response, subjects were instructed to press a button on a computer-controlled response box to indicate completion of the trial. After all of the subjects had responded, the next trial was initiated.

In the MRT, subjects identified 100 consonant-vowel-consonant words, presented one at a time. After hearing each word, subjects were presented with a list of six alternative words centered on a CRT. The subjects were instructed to press one of six buttons on a computer-controlled response box to indicate which word was heard. All subjects were instructed to respond as quickly and as accurately as possible. Response accuracy and latency was recorded by a computer.

After the tests of isolated word recognition, recognition performance for words in fluent sentences was measured using the Harvard psychoacoustic sentences and the Haskins anomalous sentences. After each sentence was presented, subjects wrote the words in the sentence in the order that they had heard them. Subjects were encouraged to guess if necessary. After all of the subjects indicated that they had completed their response (by pressing a button on a response box), the next trial was initiated.

Training Procedures. Except for the control group, all subjects received training with synthetic speech on Days 2 through 5. Novel-word and repeated-word subjects were trained with isolated words. Novel-sentence and repeated-sentence subjects were trained with fluent sentences. In each case, after listening to a stimulus (either a single word or a sentence), subjects transcribed the stimulus. After all of the subjects indicated that they had finished responding, feedback was provided. Feedback consisted of a visual presentation of the stimulus printed on a CRT together with a second auditory presentation of the stimulus item. After the feedback was presented, subjects pressed a button to begin the next trial and the procedure was repeated for subsequent trials.

On each day of training, novel-word subjects were presented with 100 new PB words, while the novel-sentence subjects heard 20 new Harvard sentences and 20 new Haskins sentences. On the first day of training, the repeated-word subjects also were presented with the same 100 PB words that were presented to the novel-word subjects on the first day of training. However, for the repeated-word subjects, this same list was presented again on each subsequent training day. Similarly, the repeated-sentence subjects were first presented with the same 40 sentences that were presented to the novel-sentence subjects on the their first day of training, and this list of sentences presented again on each subsequent day of training.

Results

Six subjects did not complete the experiment, and their data were excluded from statistical analyses. Of the remaining 60 subjects, 12 subjects received novel-word training, 12 received novel-sentence training, 13 received repeated-word training, 11 received repeated-sentence training, and 12 received no training.

To score a correct response, subjects had to transcribe the exact word (or homonym) with no additional or missing phonemes. For example, if the word was flew and the response was flute or few, the response was scored as incorrect. However, flue would have been scored as a correct response.

The principal results of the present study concern the differences among the five groups of subjects on Day 1 (the pre-training test session) and Day 6 (the post-training test session). The analysis for each type of task will be presented separately. Performance on isolated words will be discussed first and then the data from the sentences will be described.

MRT Words. The MRT was only presented on Days 1 and 6. Thus, all groups were equally familiar with this task. Mean percent correct performance on the MRT for the five groups is presented in Table 1. An analysis of variance indicated that there were significant differences between subject groups as a result of different types of training ($F(4,50) = 3.52$, $MS_e = .00346$, $p < .01$) and performance improved significantly from the pre-training test to the post-training test ($F(1,50) = 97.8$, $MS_e = .00159$, $p < .0001$). Furthermore, the interaction between test session and training group was also significant ($F(4,50) = 6.03$, $MS_e = .00159$, $p < .001$). These results indicate that although performance improved from the pre-training to the post-training test session, the degree of improvement varied as a function of the type of training received by different subject groups. A Newman-Keuls analysis of the pre-training scores indicated that there were no reliable differences in performance among the groups prior to training. However, an examination of the improvement scores (i.e., the difference between pre- and post-training test performance revealed that all of the groups receiving training improved significantly from the pre-training to the post-training session ($p < .01$). In contrast, the control group did not demonstrate any reliable evidence of improvement. Moreover, a Newman-Keuls analysis of the improvement scores indicated that each of the training groups differed significantly from the control group ($p < .05$), but not from each other. These results clearly suggest that training with either isolated words or fluent sentences produces equivalent improvements in performance on the MRT. Furthermore, these results suggest that, as long as the number of stimulus presentations are equivalent, training with a repeated set of stimuli produces as much improvement on the MRT as training with novel stimuli.

Insert Table 1 about here

PB Words. The effects of training on isolated word recognition was also examined using the open-response set PB task. The results from this task are presented in Table 2 and are quite similar to the results from the MRT. An analysis of variance indicated that there was a significant improvement in recognition performance from the pre-training to the post-training test ($F(1,55) = 546.0$, $MS_e = .00215$, $p < .0001$) and there were significant differences in recognition performance as a consequence of different types of training ($F(4,55) = 2.87$, $MS_e = .00444$, $p < .05$). Moreover, as in the MRT data, different types of training produced differing amounts of improvement from the pre-training to the post-training test ($F(4,55) = 10.04$, $MS_e = .00215, p < .0001$).

An analysis of the pre-training test scores indicated that there were no significant differences in performance among the five groups prior to training. However, a series of planned comparisons indicated that, although the recognition performance of all five groups improved significantly from the pre-training test to the post-training test session ($p < .01$), all training groups improved significantly more than the control group ($p < .01$). Moreover, the training groups did not differ reliably from one another. Thus, for transcription of PB words, training with either words or sentences, and between training with either novel or repeated exemplars stimuli produced equivalent improvements in recognition performance. Therefore, based on the results of the PB task and the MRT, one might conclude that all four training groups acquired equivalent abilities to recognize Votrax-generated speech.

Table 1

Transcription accuracy for words in the pre-training and post-training

Modified Rhyme Task (percent correct)

	Test Session		
Condition	Pre-training	Post-training	Change(% points)
Control	65.5	66.6	1.1
Novel Words	64.6	75.8	11.2
Repeated Words	65.1	75.6	10.5
Novel Sentences	67.4	76.4	9.0
Repeated Sentences	64.9	71.3	6.4

However, the sentence perception data, described next, strongly argue against this conclusion.

Insert Table 2 about here

Harvard Sentences. The third set of stimulus materials consisted of coherent, meaningful sentences. Each sentence contained five key words that were scored for recognition accuracy. Table 3 displays the average percent accuracy scores for each group of subjects in the pre-training and post-training test sessions. An analysis of variance indicated no overall significant increase in performance from the pre-training to the post-training test session ($F(1,55) = 1.7$, $MS_e = 0.0068$, $p > .10$). However, there was a significant effect on performance of the type of training received by different subject groups ($F(4,55) = 7.95$, $MS_e = 0.0124$, $p < .001$). Moreover, there was a significant interaction between performance in the pre- and post-training test sessions and type of training ($F(4,55) = 14.93$, $MS_e = .0068$, $p < .001$) indicating that although there was no reliable overall change in performance from the pre-training to the post-training test, the direction of change varied significantly as a function of type of training.

Insert Table 3 about here

Mean performance for the control, word-trained, and sentence-trained groups did not differ significantly from each other on the pre-training test (Day 1). However, five days later, word recognition in the Harvard sentences was dramatically different for the different subject groups. Planned comparisons indicated that performance dropped significantly for control subjects and for subjects that were trained with novel words ($p < .05$). However, subjects trained with repeated words showed no reliable increase or decrease in performance. In contrast, subjects trained with either novel or repeated sentences demonstrated a significant improvement in word recognition accuracy ($p < .01$). A Newman-Keuls analysis of the improvement scores demonstrated that the novel- and repeated-sentence conditions each produced significantly greater improvement scores than the repeated-word, novel-word, and control conditions ($p < .01$) It is not immediately apparent why the performance of the control and novel-word trained subjects decreased from the pre-training to the post-training test.

Combined with the results from recognition of isolated words the results from recognition of words in Harvard sentences indicates that: (1) training with isolated words improves recognition of words in isolation, but does not improve recognition of words in fluent sentences; (2) training with sentences produces reliable improvements in recognition of words in isolation and in sentences, and (3) training with novel stimuli does not produce reliably different results from training with repeated stimuli. These conclusions are further supported by results obtained for word recognition in semantically anomalous sentences.

Table 2

Transcription accuracy for words in the pre-training and post-training PB-word task (percent correct)

	Test Session		
Condition	Pre-training	Post-training	Change(% points)
Control	26.8	36.0	9.2
Novel Words	25.5	47.3	21.8
Repeated Words	24.2	48.2	24.0
Novel Sentences	25.6	47.4	21.8
Repeated Sentences	25.8	48.1	22.3

Table 3

Transcription accuracy for words in the pre-training and post-training Harvard sentence task (percent correct)

	Test Session		
Condition	Pre-training	Post-training	Change(% points)
Control	49.2	38.0	- 11.2
Novel Words	44.7	34.7	- 10.0
Repeated Words	40.6	40.6	0
Novel Sentences	45.3	62.3	17.0
Repeated Sentences	44.4	58.4	14.0

Haskins Sentences. The results for recognition of words in syntactically correct, semantically anomalous sentences are summarized in Table 4. An analysis of variance indicated that type of training had a significant effect on recognition performance ($F(4,55) = 15.8$, $MSe = .01004$, $p < .0001$) and there was an overall increase in recognition performance from the pre-training test session to the post-training test session ($F(1,55) = 538.0$, $MSe = .00465$, $p < .0001$). Moreover, the effects of type of training on the change in performance from pre-training to post-training session was also significant ($F(4,55) = 17.2$, $MSe = .00465$, $p < .0001$). Further analyses indicated that although subject groups (receiving different types of training) did not differ significantly from one another in the pre-training test ($F(4,55) = 2.21$, n.s.), these groups differed significantly after training ($F(4,55) = 17.2$, $MSe = 0.00925$, $p < .0001$). Planned comparisons indicated that all subject groups improved significantly from the pre-training to the post-training test ($p < .01$). However, a Newman-Keuls analysis indicated that the improvement shown by the novel- and repeated-sentence groups was not significantly different from each other, but was significantly greater than that of the control, novel-word, and repeated-word groups ($p < .01$). Moreover, the control, novel-word, and repeated-word groups did not differ reliably after training. Thus, providing training with repeated or novel isolated words had no more effect on recognizing words in fluent, semantically anomalous sentences than providing no training whatsoever.

Insert Table 4 about here

Taken together, the results from the four different sets of stimulus materials presented in the pre- and post-training tests show that: (1) subjects who received no training showed little or no improvement in recognition performance, (2) all subjects who received training improved in recognition of novel, isolated words by about the same amount, and (3) subjects trained on isolated words did not show any improvement in recognizing words in fluent sentences compared to control subjects, while subjects trained on fluent sentences improved significantly on recognizing words in novel sentences. One account of the difference between word-trained and sentence-trained subjects is that subjects trained with isolated words may have difficulty locating the beginning and ends of words in fluent Votrax speech. In isolated-word recognition tasks, the beginning and end of each word is clearly marked by a period of silence. However, fluent connected Votrax speech does not contain physical segmentation cues to provide the listener with an indication of the location of word boundaries. If explicit acoustic word boundaries aid in word recognition by segmenting fluent natural speech into word-size units, subjects trained with isolated words might have expected and needed these boundaries to recognize the sentences produced by the Votrax system. However, subjects trained with fluent sentences of synthetic speech may have learned explicitly to do without these word boundaries and may have learned a strategy of segmenting words by recognizing them one at a time in the order by which they are produced. This strategy would provide a recognition advantage for the sentence-trained subjects compared to subjects trained with isolated words. This account suggests that the performance of word-trained subjects might improve when word boundary cues are provided. This hypothesis was evaluated by a more detailed analysis of the recognition performance in the anomalous-sentence task.

Table 4

Transcription accuracy for words in the pre-training and post-training Haskins sentence task (percent correct)

	Test Session		
Condition	Pre-training	Post-training	Change(% points)
Control	24.0	42.3	18.3
Novel Words	19.0	41.7	22.7
Repeated Words	21.3	41.7	20.4
Novel Sentences	25.6	65.0	39.4
Repeated Sentences	25.9	69.3	43.4

Each Haskins sentence is constructed with a fixed syntactic frame (i.e., "The _adjective_ _noun_ _verb-ed_ the _noun_"). Thus, each sentence contained two key words following the definite article _the_, and two key words following open class items (i.e., an adjective, noun, or verb) that varied from sentence to sentence and subjects were told explicitly about this invariant syntactic structure. In addition, the response sheets were marked for each sentence, with separate blanks for each open-class item and the word "the" for each occurrence of the definite article. If word-trained subjects had difficulty locating the beginnings of words, their recognition performance might be better for words following the definite article compared to control subjects, since the relative locations of the definite articles were known in advance and the word-trained subjects would have better isolated word recognition skills. In contrast, the performance of the word-trained and control groups should not differ on the words following an open-class item because no word boundary cues were provided.

In order to evaluate the hypothesis that word-trained subjects would recognize words following "the" more accurately than words following a open-class item, the data from the Haskins task was reanalyzed to include word position (words following "the" or an open-class item) as a factor. This reanalysis does not, in any way, change the statistical observations of the prior analysis of variance. Thus, only the main effect of word position and its interactions will be reported. The means for the different treatment combinations are shown in Table 5. An analysis of variance indicated that, in the pre-training test session, subjects performed significantly better on the words that immediately followed the definite article than on those that followed open class words (30% correct versus 16%, $F(1,55) = 136.75$, $MSe = .00479$, $p < .0001$). Type of training and word position did not interact in a reliable and systematic fashion.

Insert Table 5 about here

As in the pre-training test, post-training recognition performance on words that immediately followed the definite article was significantly better than performance on words that followed an open class item (67% versus 36%, $F(1,55) = 278.0$, $MSe = .00970$, $p < .0001$). However, in contrast to pre-training performance, test scores after training revealed a significant interaction between word position and the type of training given to each subject group ($F(4,55) = 6.78$, $MSe = .0097$, $p < .001$). An examination of the improvement scores for each combination of word position and subject group (shown in Table 6) reveals that much of the improvement demonstrated by the control group and the repeated- and novel-word trained groups was due to increased recognition performance on words following a definite article. Indeed, for these groups the improvement scores for words following an open class item were significantly lower than the improvement scores for words following "the" ($p < .01$). However, for the repeated- and novel-sentence trained groups, the improvement scores for words following an open-class item were not significantly different from the improvement scores following "the". Thus, subjects trained with novel or repeated sentences acquired a generalized ability to recognize words in novel, fluent sentences regardless of position. In contrast, subjects who either were trained with isolated words or who had received no training required additional cues for word segmentation to assist in word recognition performance.

Table 5

Transcription accuracy for words in the pre-training and post-training PB-word Task as a function of word position (percent correct)

	Test Session			
	Pre-training		Post-training	
Condition	After "the"	After open-class	After "the"	After open-class
Control	30.8	17.1	56.7	27.9
Novel Words	25.0	12.9	62.9	20.4
Repeated Words	27.7	15.0	60.8	22.7
Novel Sentences	32.9	18.3	73.3	56.7
Repeated Sentences	36.4	15.5	81.4	57.3

Note. The category After "the" contains words that were presented immediately after the word "the"; the category After open-class contains words that were presented immediately after an open-class word (see text for further details).

Insert Table 6 about here

This conclusion is supported further by Newman-Keuls analyses of the simple effects of word position, averaged across the novel/repeated manipulation. For words following an open-class item, sentence-trained subjects improved significantly more than word-trained and control subjects ($p < .01$). Moreover, the improvement demonstrated by word-trained subjects scores did not reliably differ from the improvement shown by control subjects.

The pattern of results for words following the definite article was quite different. The improvement scores for word-trained subjects were significantly higher than those of the control subjects ($p < .05$), but significantly lower than those of the sentence-trained subjects ($p < .05$). (The difference between sentence-trained and control subjects was also significant, $p < .01$.) Thus, as predicted, prior experience with isolated words aided recognition of words in fluent sentences only when the identity of the preceding word was known in advance, providing a cue to word boundaries. Furthermore, subjects trained with isolated words were able to recognize words in sentences more accurately than control subjects when some location information was provided. Thus, the isolated-word training only improved recognition of isolated or segmented word patterns.

These data from the anomalous sentences suggest that word-trained subjects were not able to separate their perception of the acoustic-phonetic structure of a preceding word from that of a subsequent word, except when they had prior knowledge and reliable information about the identity of the preceding word (see Nakatani & Dukes, 1977; Nusbaum & Pisoni, 1985). With that one exception, training listeners to recognize words in fluent sentences required specific experience with fluent connected speech. These results demonstrate that improvements in recognizing isolated words do not necessarily predict improvements in recognizing words in fluent sentences. Thus, the present findings argue against the hypothesis that word segmentation is a direct result of word recognition. In fact, it is possible that the skill that sentence-trained subjects acquired over and above the skills acquired by the word-trained subjects may involve explicitly learning the strategy of segmenting fluent speech via word recognition.

Although word-trained subjects were not able to generalize their newly acquired skills to recognizing words in fluent sentences, sentence-trained subjects did improve at recognizing isolated words. This suggests that the perceptual skills acquired in learning to recognize words in fluent sentences form a functional superset of the skills acquired during training with isolated words. It is interesting to compare this finding with those reported by Kolers and Magee (1978). In the Kolers and Magee study, training subjects to recognize inverted letters had little transfer to reading inverted words, and training with inverted words had little impact on naming inverted letters. Thus, Kolers and Magee found little evidence for transfer even though visual words are structurally a superset of letters (just as auditory sentences are comprised of words). The difference in the studies may arise from the differences between auditory and visual modalities. Printed letters are discrete stimuli, segmented from one another by blank spaces; in contrast, there are no silent intervals for one auditory word from its neighbors and coarticulation effects may span word boundaries. One implication is that

Table 6

Improvement (change in percentage points) from the pre-training test to the post-training test as a function of word position

	Improvement from pre-training to post-training	
	Word position	
Condition	After "the"	After open-class
Control	25.9	10.8
Novel Words	37.9	7.5
Repeated Words	33.1	7.7
Novel Sentences	40.4	38.4
Repeated Sentences	45.0	41.8

Note. _After "the"_ classifies words that were presented immediately after the word "the"; _After open-class_ classifies words that were presented immediately after an open-class word (see text for further details).

conclusions about perceptual learning of orthography cannot be generally applied, without great caution, to the domain of speech perception, and vice versa (see Liberman et al., 1967).

Despite variations in type of task (closed-response set procedures vs. open-response set procedures), type of stimulus materials (sentences vs. words), and type of training (sentences vs. words), no significant differences were observed in performance between perceptual learning of novel and repeated training sets. This result differs from the findings of Posner and Keele (1968) and Dukes and Bevan (1967), but is consistent with the findings of Nagata (1976) and Palermo and Parish (1971). One account of the present results is based on the observation that the variations in synthetic speech that must be learned are lawful and rule-goverened like the variations in artificial-language materials. Under such conditions, perceptual learning of a repeated training set appears to produce the same level of generalization learning as training with novel stimuli, as long as the number of presentations is the same for the two training procedures. However, the utility of the repeated training set for generalization learning may depend on the degree to which the repeated stimuli characterize the underlying structure of the entire ensemble of possible stimuli (Palermo & Parish, 1971). In this context, it is useful to recall that Posner and Keele found that subjects trained with a highly variable set of highly distorted exemplars, classified new, highly distorted exemplars more accurately than subjects trained with a less variable, less distorted set. Considered together, these studies suggest that the structural relationship between the training and test stimuli is far more important for perceptual learning than simply the relative number of novel stimuli presented during training.

Training Data. Additional support for the proposal that the novel and repeated stimuli both provided a sufficient characterization of the rule structures of the text-to-speech system is revealed by an examination of the data from each of the training sessions for the three types of stimulus materials -- PB words (see Figure 1), Harvard sentences (see Figure 2), and Haskins sentences (see Figure 3). These data from the training sessions shows systematic improvements for subjects trained with repeated and novel words (Figure 1), and for subjects trained with repeated and novel sentences (Figures 2 and 3). The fact that subjects trained with repeated stimuli did not reach asymptotic performance after a single training session and continued to improve throughout training, indicates that the structural complexity of the repeated stimuli could not be completely learned in a short period of time. While it is not surprising that subjects trained with novel stimuli continued to learn new attributes of these stimuli and thereby show systematic improvements in performance, it is interesting that subjects trained on repeated stimuli did not completely master these stimuli immediately and also continued to learn new information from these stimuli throughout training.

Insert Figures 1, 2, and 3 about here

For all three types of stimulus materials, recognition performance of the repeated-stimulus groups was significantly higher than performance of the novel stimulus groups (PB words, $F(1,23) = 49.6$, $MSe = .0129$, $p < .0001$; Harvard sentences, $F(1,21) = 52.6$, $MSe = .0202$, $p < .0001$; and Haskins sentences, $F(1,21) = 17.2$, $MSe = .0195$, $p < .001$). By itself, this is not a

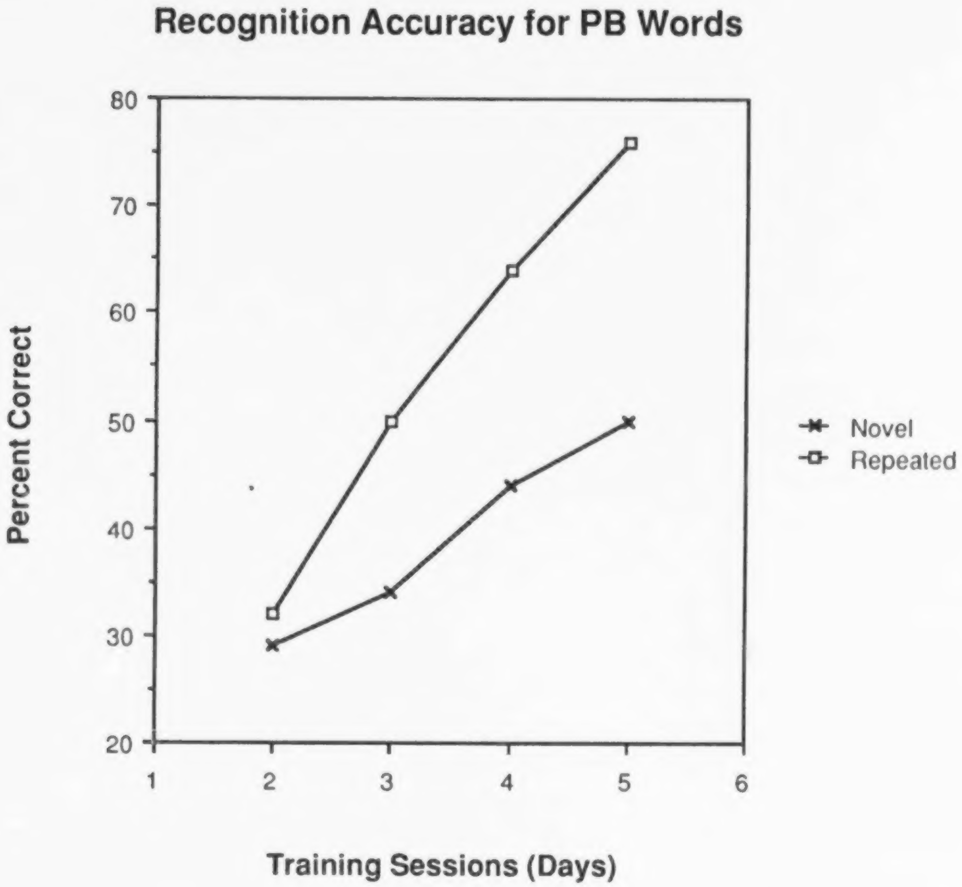


Figure 1. Transcription accuracy for words in the phonetically-balanced word lists for each day of training in Experiment 1. (Note: Days 1 and 6 were pre- and post-training test sessions.)

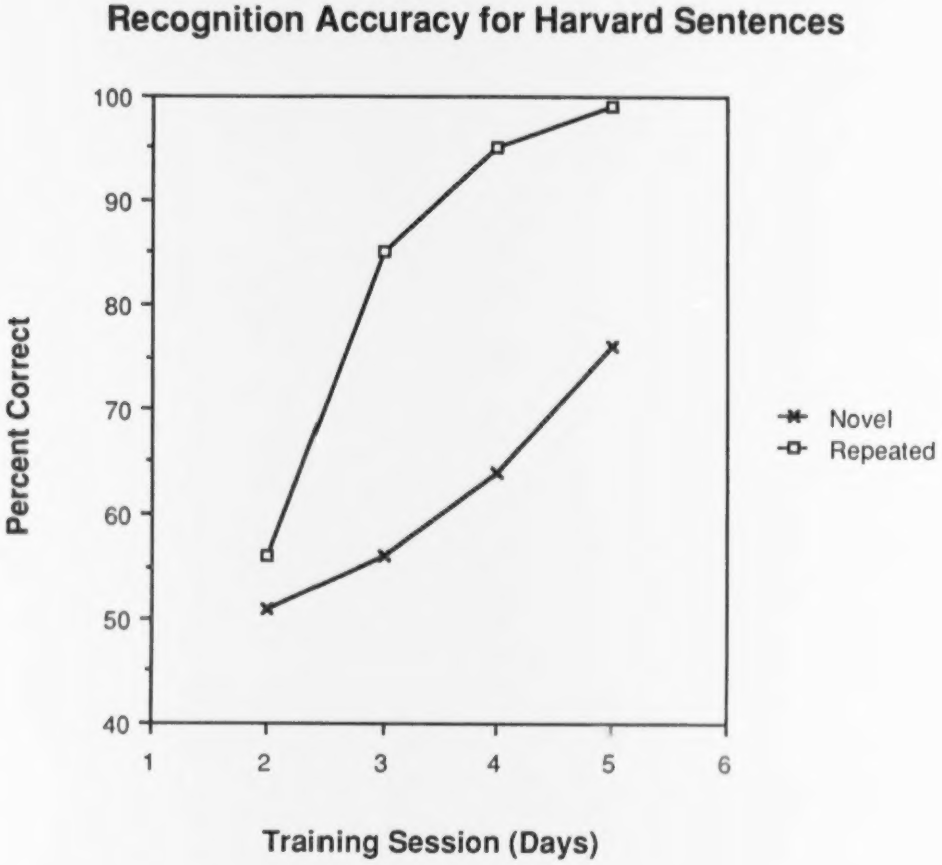


Figure 2. Transcription accuracy for words in the Harvard sentences for each day of training in Experiment 1. (Note: Days 1 and 6 were pre- and post-training test sessions.)

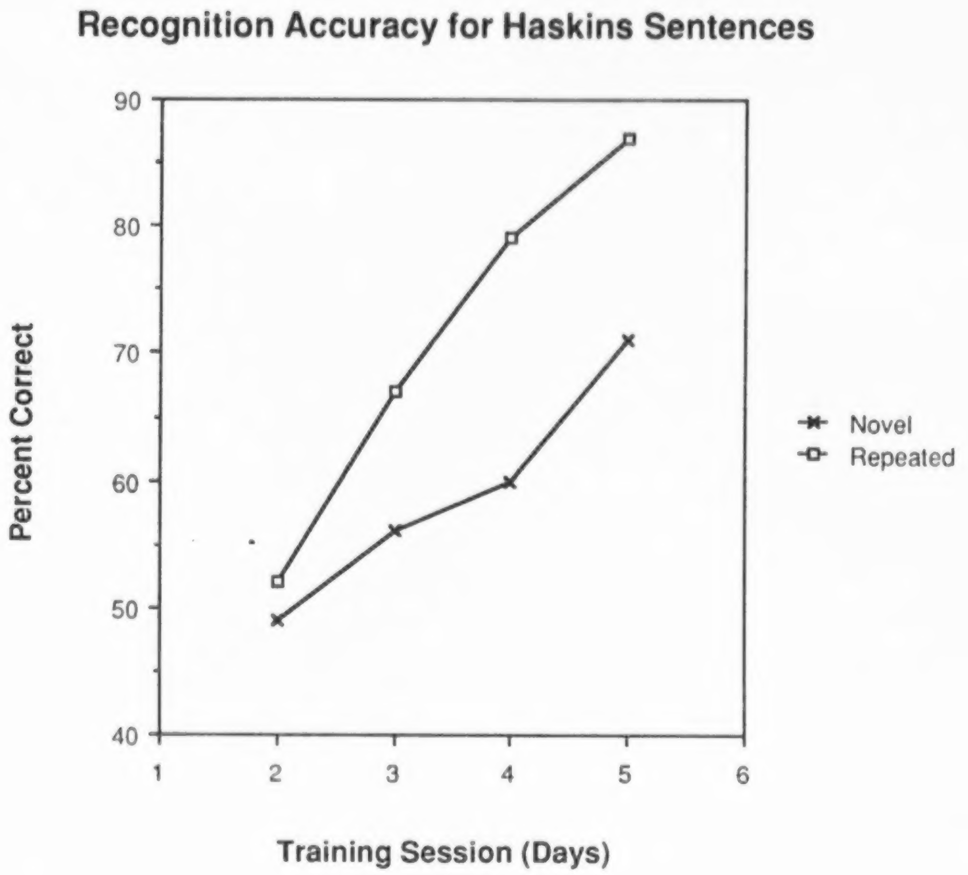


Figure 3. Transcription accuracy for words in the Hakins sentences for each day of training in Experiment 1. (Note: Days 1 and 6 were pre- and post-training test sessions.)

surprising finding because the repeated-stimulus groups always recognized and received feedback about precisely the same stimuli on every day of training, while the novel-stimulus group always engaged in generalization -- they never were presented with the same stimuli twice.

In addition, overall recognition performance improved significantly on each day of training for all stimulus materials (PB words, $F(3,69) = 279.7$, $MSe = .0018$, $p < .0001$; Harvard sentences, $F(3,63) = 214.6$, $MSe = .00228$, $p < .0001$; Haskins sentences, $F(3,63) = 141.6$, $MSe = .00233$, $p < .0001$). Thus, there is reliable evidence of perceptual learning occuring throughout the training sessions. However, of greatest interest is the degree to which the amount of learning in each training session depended on the type of training provided (repeated vs. novel stimuli). In fact, the interaction between performance in each training session and type of training was significant for all three types of stimulus materials (PB words, $F(3,69) = 35.5$, $MSe = .0018$, $p < .0001$; Harvard sentences, $F(3,63) = 34.3$, $MSe = .00228$, $p < .0001$; Haskins sentences, $F(3,63) = 11.6$, $MSe = .00233$, $p < .0001$).

For all types of stimulus materials (i.e., PB words, and Harvard and Haskins sentences), paired comparisons indicated that the interaction between the type of training (with repeated or novel stimuli) and training session was due to the absence of any significant difference between the two types of training on the first day of training (Day 2 of the experiment), followed by significantly better recognition performance for repetition training for all subsequent days of training ($p < .01$). The reason for the advantage of the repetition groups over the novel-stimuli groups during the training sessions is clear: The novel-stimulus groups were always presented with new stimuli on each training day, while the repetition groups responded to the exact same stimuli on every day. It is not surprising therefore, that subjects performed better on stimuli they had prior experience with than on completely new stimuli. However, Newman-Keuls analyses of the simple effects of repetition revealed that in the repeated stimulus condition, significant improvements occurred in word recognition in repeated PB word lists and Haskins sentences ($p < .05$), and except for the last day, regular, significant improvements occurred in word recognition in the repeated Harvard sentences ($p < .05$; failure to achieve significant improvement from the third to the last day of training was probably due to a ceiling effect). This indicates that subjects in the repetition groups continued to extract new information and learn about the repeated stimuli on each subsequent training day.

In the novel stimulus conditions, significant day-by-day improvements were observed for word recognition in the PB word lists and Harvard sentences after the second day of training ($p < .05$); significant improvements in word recognition in Haskins sentences occurred from the first to the second day of training, and from the third to the last day of training ($p < .05$). The difference in the learning curves for the repeated-stimulus and novel-stimulus conditions was undoubtably due to the fact that in the repeated-stimulus condition the same list of stimuli was presented on each day of training, while in the novel-stimulus condition, subjects were always required to identify novel stimuli. This demonstrates that repeated presentation of a word will increase the probability that it will be identified in a subsequent presentation. Identifying new words is simply more difficult than recognizing a repetition of a previously identified word. Notably, however, performance in the repeated stimuli condition generally did not reach asymptotic levels.

Discussion

The results of the present experiment replicate and substantially extend the findings reported by Schwab et al. (1985) for perceptual learning of synthetic speech generated by rule. First, all subjects who received training with synthetic speech displayed significant perceptual learning. All subjects performed better in recognizing isolated words produced by a text-to-speech system after training than before training. Moreover, this learning was measured by performance in recognizing completely novel words. Thus, subjects clearly learned an abstract representation of the acoustic-phonetic properties of Votrax-generated synthetic speech that significantly aided in word recognition.

However, although all subjects demonstrated reliable generalization learning in identifying novel stimuli after training, a second major finding concerned generalization learning across stimulus materials. Subjects trained on fluent sentences improved in recognition performance for both isolated words and words in sentences. In comparison, subjects trained on isolated words improved only in recognition of isolated words. In general, with one exception, word-trained subjects showed no better performance for recognizing words in sentences than subjects who had received no training at all. The exception was that when subjects were given some information about the location of a word in a sentence (i.e., words following "the" in anomalous sentences), word recognition performance for word-trained subjects was better than for control subjects.

Considering current theories of auditory word recognition, these findings are somewhat surprising since these theories posit that improvements in recognition of isolated words should convey the same advantage for recognition of words in sentences. The prediction of these theories should be especially true for the speech generated by the Votrax system because the sentences produced by this system are just concatenated strings of words -- there are no sentence-level phenomena in this synthetic speech. Contrary to most theories of word recognition, the present findings suggest that sentence-trained subjects learned something about sentences that could not be learned from training with isolated words alone. One hypothesis is that sentence-trained subjects learned to recognize words in sentences without explicit segmentation cues. A corollary of this hypothesis is that word-trained subjects performed poorly at recognizing words in sentences because they expected the type of word boundary phenomena that normally occur in natural speech. The absence of these phenomena dictates the need for learning a strategy that most theories of word recognition attribute to the listener as part of normative word recognition -- segmentation by recognition. Perhaps it is this strategy that is learned by subjects trained with fluent sentences.

Another major finding of the present study was that subjects trained with the same stimuli every day showed as much generalization learning as subjects trained with novel stimuli. Therefore, it is clear that generalization learning is not dependent on training with novel stimuli. This is, in some respects, a very surprising finding because subjects trained with novel stimuli everyday were continually engaged in generalization. Subjects trained with repeated stimuli did not engage in generalization until the final session of the study. As a consequence, we might expect subjects with more experience at generalization to perform better on a generalization task, while subjects trained on a fixed set of stimuli should perform much better on those stimuli but show little generalization to completely new stimuli.

74

One indication of the reason for this outcome may be found in the training data. In general, repeated-stimulus and novel-stimulus groups continued to improve in performance throughout the training sessions without reaching an asymptote in accuracy. Remember that these data come from two very different sets of stimuli. For the subjects trained on novel stimuli each day, the training data reflect day-by-day generalization performance. However, for the subjects trained on repeated stimuli, the training data reflect increments in recognition performance for the same stimuli on each day. Thus, it is clear that subjects did not quickly or easily master the training set even though it was presented on each day with feedback. The apparent difficulty in learning this repeated training set may reflect the degree to which the training set characterizes the rules used by the Votrax in synthesizing speech. As in the research on artificial-language learning, if the training set sufficiently describes the actions of the rules, generalization learning can occur even if the training set is relatively restricted.

Alternatively, the equivalent effectiveness of training with repeated and novel stimuli could be due simply to the number of training stimuli presented rather than the structural complexity of the training set. Since the subjects in repeated- and novel-stimulus training groups had the same exposure to synthetic speech and the same amount of feedback, it is possible that generalization learning in this experiment was strictly due to the number of stimulus presentations during training for each group. In Experiment 2, we tested this hypothesis by presenting two groups of subjects with the same number of stimuli during training, but we substantially reduced the structural complexity of the training set for one of the groups.

Experiment 2

Experiment 2 was designed to examine further the difference between training with repeated and novel exemplars. In Experiment 1, no systematic differences were observed between perceptual learning with repeated and novel stimuli despite variations in tasks and type of stimulus materials. Of greatest interest is the finding that repeated- and novel-stimulus training produced equivalent patterns of generalization learning. Based on previous research on artificial-language learning, it is tempting to conclude that the set of stimuli presented in the repeated stimulus condition was complex and varied enough to provide a reasonable characterization of the underlying structure of the entire ensemble of synthetic speech generated by the Votrax text-to-speech system. Learning the acoustic-phonetic structure of the synthetic speech may have been aided by prior knowledge of the lexical structures of English. Thus, while subjects in previous studies learned highly arbitrary and novel stimulus-response mappings, the subjects in Experiment 1 learned to map a "distorted" set of acoustic-phonetic cues onto a previously well-learned set of relations among natural acoustic-phonetic cues and lexical knowledge. Subjects may have modified existing knowledge structures to incorporate new acoustic-phonetic representations.

An alternative account of the perceptual learning observed in the first experiment is that simple exposure to the mechanical "sound" or voice quality of the Votrax-generated speech may improve the intelligibilty of the speech. According to this hypothesis, there should be no difference between repeated and novel stimulus conditions as long as the amount of exposure to the synthetic speech is the same in repeated- and novel-stimulus training conditions. To investigate this hypothesis, we trained subjects on 200 novel

PB words or on a fixed set of 10 PB words that was repeated 20 times. Such a small set of repeated stimuli is unlikely to provide a reasonable characterization of the underlying acoustic-phonetic rules of the text-to-speech system and is also very likely to be learned completely with a small number of repetitions. Thus, we carried out this second experiment to determine whether the generalization learning in the first experiment was strictly a result of the equivalent exposure to synthetic speech for the repeated-stimulus and novel-stimulus groups. In this experiment, if repeated- and novel-word groups display equivalent generalization learning, that would provide evidence that perceptual learning of synthetic speech is a direct function of exposure regardless of the structure of the training set. On the other hand, better generalization learning by the novel-word group relative to the repeated-word groups would provide evidence that effective generalization of perceptual learning depends on the structural properties of the set of training stimuli.

Method

Subjects. Seventy-two undergraduates participated in this experiment to fulfill a requirement for an introductory Psychology course. All subjects reported that English was their first language, that they had no history of hearing or speech disorders, and that they had no prior exposure to synthetic speech.

Design. Subjects were assigned to two groups of thirty-six, and all subjects completed the experiment. Both groups were given a pre-training, open-response test of 50 PB words at the beginning of a one-hour session, and both received a post-training open-response test of 50 PB words at the end of the session. However, in the interval between the pre-training and post-training tests, the two groups were trained with different types of materials. One group was trained with 200 novel PB words divided into 4 blocks. The other group received a fixed list of 10 PB words. This list was repeated 20 times during training.

Materials. The PB word lists used in Experiment 1 were modified for this experiment. Four of the eight 50-word lists used for training in Experiment 1 were used without modification for training subjects in the novel-word condition. The list of 10 words that was used to train subjects in the repeated-word condition was constructed from the 100 PB words used during the first day of training in Experiment 1. These words were selected because, in Experiment 1, they were difficult to identify on the first day of training, but were very reliably identified by the repeated-word subjects on the last day of training. Thus, although these words were difficult to identify initially, they were easily learned in the first experiment. Although all 10 words were always presented in each set of 20 trials, their order within a block of trials was randomly varied.

The pre-training and post-training test lists each contained 40 novel PB words, plus the 10 words that were used to train subjects in the repeated stimulus condition of the present experiment. These sublists will be referred to as the 40-word subtest and the 10-word subtest, respectively. The words of the 10-word subtest were randomly mixed with those of the 40-word subtest in both the pre-training and post-training stimulus lists.

Procedure. The same apparatus and general procedures used in Experiment 1 were also used in the present experiment. Subjects were tested in groups, with a maximum of six subjects per group. Subjects were told that they would be listening to single monosyllablic words produced by a text-to-speech system. For the pre-training test, they were instructed to listen carefully to each word and after each word was presented, to write the word that they heard. Subjects were told to guess if they were uncertain about a word's identity. For the training sessions, subjects were told that after transcribing their response, they would be shown the actual word on a CRT monitor along with a second auditory repetition of the word. All subjects were told that words could be repeated within a list. The post-training test procedure was identical to that used in pre-training test (i.e., no feedback was provided to subjects). In all other respects, the training and testing procedures used in the present experiment were identical to those used in Experiment 1.

Results

Transciption accuracy was determined according to the procedure used in Experiment 1 for PB words. The principle results concern the performance of the novel- and repeated-stimulus groups during the test sessions. However, because the logic of Experiment 2 requires that the repeated word list is completely learned prior to the post-training test, the results for the training session will be reported first.

Training Data. Performance during training on the novel- and repeated-word lists is summarized in Figure 4. To facilitate comparison with the 4 blocks of 50 novel words, the 20 repetitions of the 10-word list were grouped into 4 blocks, each containing 5 repetitions of the 10 words. All of the subjects in the repeated-stimulus condition acheived perfect performance by the third block of trials. An analysis of variance of the training data for the subjects in the novel-word condition indicated that performance increased significantly from the first training list to the last training list ($F(3,105) = 52.4$, $MS_e = .00273$, $p < .0001$). A Newman-Keuls analysis indicated that percent correct word recognition increased from the first training list to the second, and from the third training list to the fourth. These data are not surprising and serve to demonstrate that subjects in the novel-stimulus condition continued to improve in recognition performance throughout training, while those subjects in the repeated-stimulus condition reached perfect performance by the third block of trials.

Insert Figure 4 about here

The superior performance of the subjects trained with repeated words on those words was apparent in performance on their first ten trials of training. Repeated-word subjects were able to correctly identify 33% of the first ten training words they received, while the novel-word subjects were able to identify only 23% of their first ten words ($F(1,70) = 8.38$, $MS_e = .0215$, $p < .01$). Thus, words produced by the Votrax text-to-speech system were more accurately identified after a single prior presentation in the pre-training test. Notably, this repetition effect occurred in the absence of any feedback

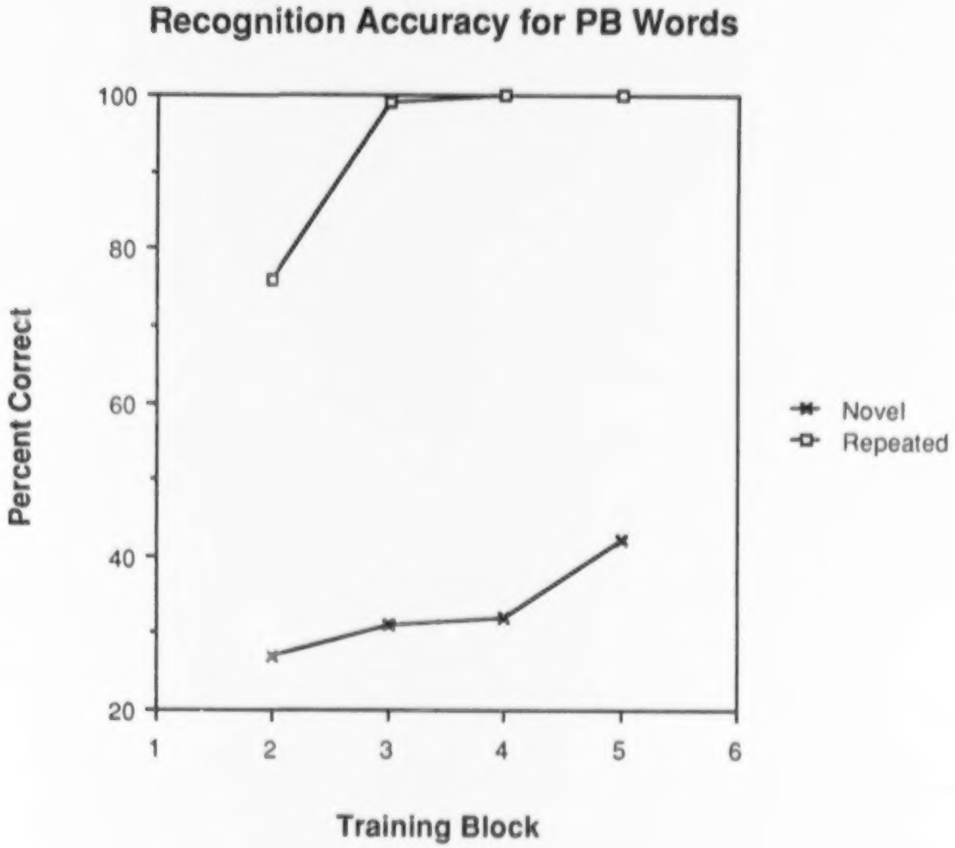


Figure 4. Transcription accuracy for words in the phonetically-balanced word lists for each training list in Experiment 2. (Note: Lists 1 and 6 were pre- and post-training test lists.)

during the initial presentation of the words. Thus, the repetition effect reported by Jacoby (1983) for visual stimuli was found in the present experiment for synthetic speech.

Testing Data. The results for the pre- and post-training tests are summarized in Table 7. Prior to training, there was a significant difference in performance on the two subtests such that subjects were able to correctly identify only 15% of the 10-word subtest compared to 20% correct responses on the 40-word subtest ($F(1,70) = 14.55$, $MSe = .00644$, $p < .001$). This difference was probably due to the fact that the 10-word subtest was constructed to be difficult prior to training, but easy after a moderate amount of training. In addition, prior to training, there was no performance difference between subject groups and no interaction between subject groups and subtest ($p > .25$).

Insert Table 7 about here

However, after training, the two subject groups performed quite differently. In order to describe these differences and to facilitate comparison with the pre-training test scores, two analyses were conducted: one on the 10-word subtest results and one on the 40-word subtest results. In both cases, subject group (novel- vs. repeated-word training) was the between-groups factor and test list (pre-training vs. post-training) was the within-groups factor.

An analysis of variance of the 10-word subtests indicated that repeated-word subjects identified 57% of the 10-word list correctly and that this was significantly greater than the 18% correctly identified by the novel-word subjects ($F(1,70) = 310.6$, $MSe = .01716$, $p < .0001$). Furthermore, significantly more items were correctly identified in the post-training subtest (60%) than in the pre-training subtest (15%), ($F(1,70) = 855.0$, $MSe = .00847$, $p < .0001$). Also, the interaction between these types of training and test list was significant ($F(1,70) = 541.3$, $MSe = .00847$, $p < .0001$). Paired comparisons indicated that although no significant difference between subject groups was found in the pre-training 10-word subtest, a significant difference between groups was observed for the post-training 10-word subtest ($p < .01$). Furthermore, although both groups showed significant improvements in performance from the pre-training 10-word subtest to the post-training subtest ($p < .01$), the improvement demonstrated by the repeated-word subjects was significantly greater than that shown by the novel-word subjects ($p < .01$). These results demonstrate that the subjects in the repeated-word condition learned to recognize words in their 10-word training list much more accurately than subjects who did not receive those words during training.

The primary concern of the present experiment is with the amount of generalization learning demonstrated by the two subject groups by recognition performance for novel words after training. An analysis of variance of the 40-word subtest indicated that more words were correctly identified in the post-training subtest than in the pre-training subtest (27% and 20%, respectively), ($F1,70) = 79.9$, $MSe = .0023$, $p < .0001$); but there was no significant overall difference between subject groups ($F(1,70) = 2.3$, $MSe = .00434$, $p > .10$). However, the difference between the pre- and post-training 40-word subtests was significant for the the two subject groups ($F(1,70) =$

Table 7

Transcription accuracy for words in the pre-training and post-training tests (percent correct)

	Test Session			
	Pre-training		Post-training	
Condition	10-word subtest	40-word subtest	10-word subtest	40-word subtest
Repeated	16.7	20.7	97.2	25.6
Novel	13.9	20.1	23.1	29.5

8.21, $MSe = .0023$, $p < .01$). A series of paired comparisons probing this interaction revealed that, although no reliable difference was observed between the two subject groups in the pre-training 40-word subtest, novel-word training produced significantly better generalization performance than repeated-word training in the post-training 40-word subtest ($p < .01$). Moreover, a comparison between pre-training and post-training results demonstrated that novel-word subjects improved significantly more than repeated-subjects as a result of training ($p < .01$). In short, subjects trained with novel words displayed significantly greater generalization learning than subjects trained with a repeated set of easily learned words, even though the number of stimuli presented to each group of subjects was equivalent.

Discussion

The findings of the present experiment clarify and extend the results of Experiment 1. The purpose of this second experiment was to determine whether simple repetition is sufficient to produce equivalent generalization learning to training with novel stimuli. One major difference between these experiments was in the structure of the repeated training set. In the first experiment, the training set consisted of a large number of words or sentences that represent a great deal of the segmental variability that occurs in speech produced by the text-to-speech system. In contrast, in the present experiment, subjects trained with repeated words were exposed to a substantially smaller set of exemplars of the Votrax generated speech. Furthermore, we chose the words for the repeated training set based on the fact that in the first experiment, these words were initially difficult to recognize but subjects were able to learn them after some training. The results from the training blocks for the repeated words demonstrates that subjects in the repeated-word group quickly learned to recognize these items perfectly. By comparison, recognition performance of the novel-word group was substantially poorer during the training blocks.

However, despite significantly poorer performance during training, the novel-word trained subjects showed significantly better performance on the generalization test than subjects trained with repeated words. Indeed, a comparison of performance on the post-training test for the 10-word and 40-word subtests indicates a large interaction between training and performance on these tests: Subjects trained with repeated words recognized those words more accurately after training than subjects trained with novel words; subjects trained with novel words recognized entirely new words in a generalization test more accurately than subjects trained with repeated words. These results demonstrate clearly that repetition alone is insufficient to facilitate generalization learning. Even though both groups of subjects had the same amount of exposure to the synthetic speech, the novel-word group showed significantly better generalization performance.

There are two conclusions that can be drawn from this pattern of results. First, it is clear that simple exposure to the mechanical sound of synthetic speech is insufficient to facilitate generalization learning. Repetition alone will not produce generalized perceptual learning. Second, by extension from the first experiment, generalization learning depends on exposure to a training set that sufficiently characterizes the rule structure of the speech. In other words, continuous presentation of novel stimuli during training is not necessary to produce generalization training. Thus, generalized perceptual learning of synthetic speech is a consequence of sampling the space of possible stimuli in such a way as to describe the structural properties of the speech.

In general, the results of the present experiment using synthetic speech produced by rule are similar to the pattern of results reported by Dukes and Bevan (1967) for naming pictures. However, one exception was observed. In the Dukes and Bevan study, subjects trained with novel stimuli did better than repeated-stimuli subjects on the new test stimuli, and subjects trained with repeated-stimuli did better than novel-stimulus subjects on old test items. However, Dukes and Bevan found that even the novel-stimulus subjects identified old items more accurately than novel items. In contrast, in the present experiment, the improvement of the novel-word subjects on the 10-word subtest was not significantly different from their improvement on the 40-word subtest. Thus, there was no distinction in recognition performance for new items and old items for novel-word subjects, even though this difference was obtained for the repeated-word subjects. This finding suggests that, in the present experiment, the old items were either not sufficiently distinctive to retain a salient episodic to further aid recognition (cf. Jacoby, 1983) or the old items were not really "learned" as individual items but were only learned for their acoustic-phonetic structure. This would mean that in learning to recognize these words, subjects were actually attending to the acoustic-phonetic properties of the speech rather than to the words themselves.

General Discussion

The present research has been concerned with the influence of stimulus repetition, stimulus variablity and stimulus structure on generalization learning. The results of this research have replicated a previous study by Schwab et al. (1985) demonstrating that listeners can learn to recognize synthetic speech more accurately with modest amounts of training. In fact, subjects are not simply learning to recognize word patterns that they have been trained on, but instead, they learn to recognize words and sentences they have never heard before. The purpose of the present study was to investigate this generalization learning in greater detail. In the first experiment, we found that training with isolated words is not equivalent to training with fluent sentences. Subjects trained with fluent sentences displayed two types of generalization learning -- they improved in recognizing words in novel sentences and they also improved in recognizing isolated words. Subjects trained with isolated words only improved at recognizing novel isolated words and did not show any better performance for recognition of words in sentences than control subjects.

Another important finding of the first experiment was that subjects trained with repeated stimuli displayed equivalent generalization learning to subjects trained with novel stimuli. Thus, subjects who engaged in generalization throughout training were no better at this task after this training than subjects who were trained on a repeated stimuli. An examination of the day-by-day performance of these two groups of subjects indicated that both groups showed systematic improvements on each training day indicating that the repeated-stimulus subjects continued to learn new information over the course of training, as did the novel-stimulus subjects. By comparison, in the second experiment, when repeated-stimulus subjects were able to quickly and completely learn the fixed training set, the novel-stimulus subjects performed significantly better on the generalization test.

In order to consider the implications of the present results for perceptual learning in general, it is useful to compare the structural characteristics of the stimulus materials used in the present experiments with those used in previous perceptual learning studies. The variability of the acoustic-phonetic structure of synthetic speech generated by the Votrax text-to-speech system is lawful and context-conditioned; moreover, this acoustic-phonetic structure, in principle, is systematically related to the acoustic-phonetic structure of English (although there are significant differences; see Yuchtman et al., 1985). Thus, subjects in the present study were faced with the problem of mapping a distorted, but systematic, set of acoustic-phonetic cues onto a previously well-learned set of relations between natural acoustic-phonetic cues and lexical representations in memory. The speech produced by a text-to-speech system is governed by a set of rules that describe the use of a particular phoneme or allophone in a specific context. However, the acoustic-phonetic structure of synthetic speech does not incorporate all the rich and redundant context-conditioned variability that represents natural speech. Instead, the acoustic-phonetic structure of synthetic speech is constrained much more severely and is limited to a small, fixed inventory of sounds. Thus, in learning to recognize Votrax-generated words and sentences, listeners are really learning to map the limited sound inventory of the Votrax speech onto already well-known phonetic categories and then they must also learn to recognize sequences of these segments as words. Because the inventory of sounds produced by the Votrax is quite limited and are systematically related to each other through acoustic-phonetic and phonological constraints, listeners may have been able to learn the acoustic-phonetic structure of the synthetic speech from a relatively small set of repeated exemplars in the first experiment.

The stimuli used by Posner and Keele (1968) were quite different: The exemplars were stochastic distortions of four dot patterns that served as prototypes. Thus, the relationship among the Posner and Keele stimuli was not as well-defined as the relationship among the present stimuli. The stimuli used by Bevan and Dukes (1967) were perhaps more systematically related (they used pictures of individuals in different poses), but it is unlikely that even these stimuli provide the rich structural coherence and redundancy of speech. The present results suggest that the underlying structure defining a set of exemplars can greatly effect the outcome of a perceptual learning experiment. In the previous two studies, as well as those conducted by Nagata (1976) and Palermo and Parrish (1971), stimulus variability substantially affected generalization learning. In the present experiment, stimulus variability had a negligible effect, as long as the stimulus set in the repeated-stimulus condition was large enough to provide a reasonable characterization of the range or perceptual space of the stimuli generated by the Votrax text-to-speech system.

In interpreting the present results, it is important to note that mere familiarity with the mechanical sound of Votrax was not sufficient to improve intelligibility. Listeners are clearly not simply becoming accustomed to the unusual sound of synthetic speech or to the sound of Votrax speech, in particular. Rather, listeners are learning very specific information about the structural properties of the speech that is produced by the rule system.

Further support for the conclusion that listeners are learning specific structural properties of the synthetic speech produced by the Votrax system comes from the results of the first experiment comparing perceptual learning for subjects trained on isolated words and subjects trained on fluent sentences. Although both groups of subjects recognizes isolated words more accurately following training, it is apparent that training with sentences did not display equivalent effects to training on isolated words. Subjects who were trained on fluent sentences also showed performance improvements for recognition of words in sentences, while subjects trained on isolated words did not show this improvement for recognition of words in sentences.

This finding is interesting for two reasons. First, these results are quite different from those obtained by Kolers and Magee (1978), who found that training subjects to read inverted letters did not improve with inverted words and vice versa. The difference in the findings obtained in these two studies is probably most directly the result of differences in the nature of the patterns learned by subjects for synthetic speech and inverted print. In the Kolers and Magee study, subjects may have adopted entirely different perceptual organizations for patterns depending on the nature of their training. Word-trained subjects may have learned to code entire word patterns while letter-trained subjects may have adopted a strategy of directing attention at subword patterns. By comparison, in the present study, word-trained and sentence-trained subjects both clearly displayed better recognition of isolated words after training. This indicates that, after training, subjects were more accurate in mapping sound sequences produced by Votrax onto the intended lexical representations. Despite the differences in training materials, both groups of subjects improved in word recognition performance. This indicates that sentence-trained subjects did not learn to treat fluent sentences of synthetic speech as holistic entities with a qualitatively different (and more complex) pattern structure than isolated words. Unlike the subjects in the Kolers and Magee study, the sentence-trained subjects obviously recognized that sentences are made up of words and so learning to recognize required learning to recognize words. However, beyond these basic improvements in recognizing words, it appears that the sentence-trained subjects learned something more about perceiving synthetic speech: They learned how to recognize words in fluent sentences, a perceptual skill that was not conferred by training with isolated words alone.

This perceptual skill might be the ability to segment fluent synthetic speech by recognizing words one at a time in the order by which they were produced. According to this segmentation strategy, the beginnings and endings of words are not located by explicit word boundary cues or information, but simply by the process of serial word recognition. Recognition of the first word in a sentence indicates the beginning of the next word and so on. Another way of expressing this is to say that the sentence-trained subjects learned to recognize words in the absence of word boundary cues that may be expected in perceiving natural speech. Conversely, the word-trained subjects, when presented with sentences, expected these boundary cues and the absence of these cues impaired word recognition performance for these subjects.

This brings up the second reason for the importance of the asymmetry of learning displayed by sentence-trained and word-trained subjects. A number of theories and models of word recognition propose that words are segmented out of fluent natural speech as a direct result of word recognition (e.g., Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986). According to these theories, fluent natural speech includes no explicit process for dividing speech into word-size units that are then matched against lexical representations. Instead, segmentation is a direct by-product of the recognition process (e.g., Reddy, 1976; Pisoni, 1978).

Suppose that these theories are correct and segmentation is a consequence of word recognition and not a necessary antecedent to it. The sentences produced by the Votrax text-to-speech system are simple concatenations of isolated words and so, in these sentences, there are no explicit acoustic cues to word boundaries as have sometimes been observed in natural speech (e.g., Nakatani & Dukes, 1977). Thus, these sentences represent precisely the type of stimuli that would be expected during word recognition by these theories. Thus, any improvements in recognizing isolated words should directly improve recognition of words in sentences as well. However, despite their improvements in isolated word recognition, the word-trained subjects did not show any better recognition of words in sentences than control subjects, with one exception. The case in which word-trained subjects performed better than the control subjects in recognizing words in sentences was for words whose beginnings were located a priori on the response sheets. In other words, when word-trained subjects knew the location of a content word, they were able to recognize it more accurately after training. This demonstrates that the major difficulty experienced by these subjects was in locating the words in the fluent speech. When the location of a word was known, these subjects could apply their enhanced word recognition skills. By extension, this suggests that sentence-trained subjects, in learning to recognize words more accurately in sentences, were really learning to recognize words in the absence of segmentation cues. However, why should subjects need to learn this if most theories of word recognition are correct and segmentation is a consequence of recognition? Similarly, if these theories are correct, why did word-trained subjects only display improved word recognition for sentences when word location information was provided? One possibility is that a fundamental claim of most theories of word recognition is wrong: Segmentation is not a consequence of the recognition process but perhaps an important antecedent or corollary of recognition.

Taken together, the pattern of results obtained in these experiments argues that listeners do not normally recognize words one at a time, in the order by which they are produced as a means of locating word boundaries. Indeed, there is increasing evidence that there may be reliable cues to word boundaries in both the acoustic structure (e.g., Nakatani and Dukes, 1977) and the phonotactic structure (Lamel and Zue, 1984) of fluent speech. Recently, Quene (1985) has shown that adding acoustic word boundary cues to synthetic speech does enhance the intelligibility of the synthetic speech. Thus, there are word boundaries in natural speech that may aid in the process of word recognition and these boundaries are absent from Votrax-generated synthetic speech. If these word boundaries do indeed play an important role in the recognition process, most of the current theories of auditory word recognition would be based upon a fundamentally incorrect assumption and would require considerable revision (Grosjean, 1985; Grosjean & Gee, in press).

One response to the claim that prosodic and segmentation cues play an important role in word recognition might be to simply incorporate these cues into extant theories of word recognition. For example, in McClelland and Elman's (1986) Trace model, word beginnings or endings would be signified by boundary cue detectors that would have a role similar to the current acoustic feature detectors except that instead of signaling phonetic information to the system, these boundary detectors would directly fire to the lexical level to indicate the start or end of a word. Although it clearly would not be difficult to add cues to a theory of word recognition, either as part of the lexical representations or as an explicit signaling mechanism, this may not be the appropriate way to incorporate this information. The addition of these cues might allow a theory to emulate human performance, but it would not be dictated on computational grounds. Including these cues would not enhance the capabilities of the theory except in terms of emulating humans. Instead, it seems that the need for segmentation cues should dictate a different approach to word recognition other than the current, strictly linear, word-by-word strategy (Grosjean & Gee, in press).

Finally, it is not clear whether repetition and novel-stimulus training in Experiment 1 would have produced equivalent effects if subjects in the repetition training condition had achieved asymptotic performance during the training sessions. If subjects extract all the information from a fixed set of stimuli so that no more overt learning occurs, and repetition training and novel-stimulus training is continued on from that point, will the two types of training be equivalent. Once learning reaches asymptotic levels in a repetition condition, the effects of repetition learning on a generalization test may level off while novel-stimulus training may continue to produce increasingly better performance on a generalization test. This is an issue that warrants investigation in future research on perceptual learning.

Clearly, the results of the present experiments demonstrate the importance of studying generalization learning for stimuli that are lawfully related to previously well-learned stimulus structures, and that are internally coherent and involve context-conditioned variability. Moreover, these results indicate that it is not always advisable to infer similarities between the processes of visual pattern recognition and speech perception. The processes that mediate perceptual learning appear to be linked directly to the type of pattern structures that are presented in different modalities and, as a consequence, perceptual learning may take different forms for different types of stimulus sets across sensory systems. It is important to begin to characterize what these differences and similarities are and how they may affect the processes of perceptual learning in order to develop more general theories of perceptual learning. Simple exemplar-based models (e.g., Jacoby, 1983) or stimulus-response associating models (e.g., Shiffrin & Scheider, 1977) may be inadequate for the task of representing the full range of complexity presented by perceptual learning in different modalities for different types of stimuli. In future work, it will be necessary to investigate systems that are capable of representing the rich structural relationships that exist among different tokens of stimuli and that represent perceptual learning as a more general process of complex skill acquisition (e.g., Grossberg, 1982; Kolers and Roediger, 1984).

References

Bartlett, F. C. (1932). *Remembering, a study in experimental and social psychology*. Cambridge: Cambridge University Press.

Cole, R. A., Rudnicky, A. I., Zue, V. W., & Reddy, R. D. (1980). Speech as patterns on paper. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Cole, R. .A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology*, 1967, *74*, 178-181.

Ebbinghaus, H. (1964). *Memory*. (H. A. Ruger & C. E. Bussenius, Trans.). NY: Dover. (Original work published 1885)

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, *58*, 955-991.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes*. Hillsdale, NJ: Erlbaum.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1984). Recognition of speech spectrograms. *Journal of the Acoustical Society of America*, *76*, 32-43.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics*, *38*, 299-310.

Grosjean, F., & Gee, J. P. (in press). Prosodic structure and spoken word recognition. *Cognition*.

Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Boston, MA: Reidel Publishing.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, *37*, 158-166.

IEEE. (1969). IEEE recommended practice for speech quality measurements (IEEE No. 297). NY: IEEE.

James, W. (1890). *Psychology*. NY: Holt.

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 21-38.

Kolers, P. A. (1976). Reading a year later. Journal of Experimental Psychology: Human Learning and Memory, 2, 554-565.

Kolers, P. A., & Magee, L. (1978). Specificity of pattern-analyzing skills in reading. Canadian Journal of Psychology, 32, 43-51.

Kolers, P. A., & Roediger, H. L., III (1984). Procedures of mind. Journal of Learning and Verbal Behavior, 23, 425-449.

Lamel, L. F., & Zue, V. W. (1984). Properties of consonant sequences within words and across word boundaries. In the Proceedings of ICASSP-84, 3, NY: IEEE Press.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C., & Cooper, F. S. (1959). Minimal rules for synthesizing speech. Journal of the Acoustical Society of America, 31, 1490-1499.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 84, 452-471.

Lisker, L. (1978). Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. Status Report on Speech Research, SR-54. New Haven, CT: Haskins Laboratories.

Markel, J. D., & Gray, A. H., Jr. (1976). Linear prediction of speech. New York: Springer-Verlag.

Marslen-Wison, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 1-86.

Nagata, H. (1976). Quantitative and qualitative analysis of experience in acquisition of a miniature artificial language. Japanese Psychological Research, 18, 174-182.

Nakatani, L. H. , & Dukes, K. D. (1977). Locus of segmental cues for word juncture. Journal of the Acoustical Society of America, 62, 714-719.

Nusbaum, H. C., Dedina, M. J., & Pisoni, D. B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. Speech Research Laboratory Technical Note 84-02. Bloomington, IN: Speech Research Laboratory, Indiana University.

Nusbaum, H. C., & Pisoni, D. B. (1985). Some constraints on the perception of synthetic speech. Behavior Research Methods and Instrumentation, 17, 253-242.

Nye, P. W., & Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Status report on speech research, SR-37/38, 169-190. New Haven, CT: Haskins Laboratories.

Palermo, D. S., & Parrish, M. (1971). Rule acquisition as a function of number and frequency of exemplars presented. Journal of Verbal Learning and Verbal Behavior, 10, 44-51.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Ed.), Handbook of learning and cognitive processes: Volume 6, Linguistic functions in cognitive theory. Hillsdale, NJ: Erlbaum.

Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. Journal of the Acoustical Society of America, 78, 381-388.

Pisoni, D. B., Aslin, R. N., Perry, A. J., Hennesy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 8, 297-314.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Quene, H. (1985). Consonant duration as a perceptual boundary cue in Dutch. In Progress Report of the Institue of Phonetics, 10, University of Utrect, The Netherlands.

Reddy, D. R. (1976). Speech recognition by machine: A review. Proceedings of the IEEE, 64, 501-531.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. Human Factors, 27, 395-408.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1368.

Strange, W., & Jenkin, J. J. (1978). Role of linguistic experience in the perception of speech. In R. D. Walk & H. L. Pick (Eds.), Perception and experience. NY: Plenum Press.

Woodworth, R. S. (1938). Experimental psychology. NY: Holt.

Yuchtman, M., Nusbaum, H. C., & Pisoni, D. B. (November, 1985). Consonant confusions and perceptual spaces for natural and synthetic speech. Presented at the 110th meeting of the Acoustical Society of America, Nashville.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Trading Relations, Acoustic Cue Integration, and Context Effects in Speech Perception*

David B. Pisoni and Paul A. Luce

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

*This is the final draft of a paper prepared for the NATO Advanced Research Workshop, "The Psychophysics of Speech Perception," held at Utrecht University, the Netherlands, June 30 - July 4, 1986. This research was supported, in part, by NIH Research Grant NS-12179 to Indiana University in Bloomington and an award to the first author from the James McKeen Cattell Fund.

Trading Relations, Acoustic Cue Integration, and Context Effects in Speech Perception

The study of speech perception differs in several very important ways from the study of general auditory perception. First, the signals typically used to study the functioning of the auditory system have been simple, discrete, and well-defined mathematically. Moreover, they typically vary along only one perceptually-relevant dimension. In contrast, speech sounds involve very complex spectral relations that typically vary quite rapidly as a function of time. Changes that occur in a single perceptual dimension almost always affect the perception of other attributes of the signal. Second, most of the basic research on auditory perception over the last four decades has been concerned with problems surrounding the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanisms. In the perception of complex sound patterns such as speech, the relevant mechanisms are, for the most part, quite centrally located. Moreover, while many experiments in auditory perception and sensory psychophysics have commonly focused on experimental tasks involving discrimination of both spectral and temporal properties of auditory signals, such tasks are often inappropriate for the study of more complex signals including speech. Indeed, in the case of speech perception and probably the perception of other complex auditory patterns, the relevant task for the observer is more nearly one of absolute identification rather than differential discrimination. Listeners almost always try to identify, on an absolute basis, a particular stretch of speech or try to assign some label or sequence of labels to a complex auditory pattern. Rarely, if ever, are listeners required to make fine discriminations that approach the limits of their sensory capacities.

Given the published literature on the perception of simple auditory signals, it is generally believed, at least among researchers in the field of speech perception, that a good deal of what we have learned from traditional auditory psychophysics using simple sinusoids is only marginally relevant to the study of speech perception. Perhaps some of what is currently known about speech perception might be relevant to the perception other complex auditory patterns which have properties that are similar to speech. At the present time, there are substantial gaps in our knowledge about the perception of complex signals which contain very rapid spectral changes such as those found in speech. And, there is little if any research on the perception of complex patterns that have the typical spectral peaks and valleys that speech signals have. Finally, our knowledge and understanding of patterns containing amplitude variations like the complex temporal patterns found in speech are also quite meager at this time. Obviously, there is a lot of basic research to do.

A voiced (periodic) speech signal is typically thought to be produced by excitation of a time-varying filter with a source spectrum which has harmonics at multiples of the fundamental. For unvoiced (aperiodic) signals, the situation is somewhat more complicated because the source spectrum is continuous and may contain energy at all frequencies and the location of the energy in the vocal tract can occur at a number of different locations between glottis and lips. However, in considering only voiced sounds, it has been convenient to assume, for modeling purposes, that the interactions between source and filter are minimal and thus it is theoretically convenient to dissociate properties related to the source spectrum from properties imposed by the vocal-tract transfer function. The relevant perceptual attributes for the perception of segmental sounds of speech are closely associated with changes in spectral shape over time. In contrast, the relevant perceptual attributes for the perception of suprasegmental or prosodic attributes of speech are related to the changes in the temporal properties of speech, such as duration, and variations in pitch and amplitude as a function of time. Considering only the segmental properties of speech sounds in patterns of word-length size, it is possible to generate an enormously rich set of highly distinctive acoustic patterns (i.e., words) that can be identified and responded to very rapidly by human listeners. When interest is directed to prosodic attributes of speech and some of the properties related to source characteristics, it immediately becomes apparent that an even richer and more distinctive set of complex signals can be generated by the combination of only a small number of variations on a larger set of perceptually-relevant dimensions.

As Pollack (1952) demonstrated over thirty years ago, speech sounds represent a class of signals that are able to transmit relatively high levels of information with only gross variations in perceptually-distinctive acoustic attributes. In other words, speech is an efficient signaling system because of its ability to exploit fundamental processing strategies of the auditory system. This theme has been taken up and expanded recently by Stevens (1980) who argues that speech signals display a certain set of general properties that set them apart from other signals in the listener's auditory environment. According to Stevens, all speech signals have three general properties or attributes in common. First, the short-term power spectrum sampled at specific points in time always has "peaks" and "valleys." That is, speech signals display up and down alternations in spectrum amplitude with frequency. These peaks in the power spectrum arise from the peaks observed in the vocal tract transfer function and correspond to the formants or vocal resonances that are so prominent in vowel and vowel-like sounds. The second general property that speech sounds display is the presence of up and down fluctuations in amplitude as a function of time. These variations in amplitude correspond to the alternation of consonants and vowels occurring in syllabic-like units roughly every 200-300 msec. Finally, the third general property that speech signals display is that the short-term spectrum changes over time. The peaks and valleys of the power spectrum change; some changes occur rapidly -- like the formant transitions of stop consonants, whereas other changes are more gradual like the formant motions of semi-vowels and diphthongs. According to Stevens (1980), speech sounds have these three general attributes and other sounds do not, and it is these attributes that distinguish speech sounds from other complex nonspeech sounds.

It should also be mentioned here that in addition to some of the differences in the signal characteristics between speech and nonspeech noted above, there are also very marked differences in the manner in which speech and nonspeech signals are processed (i.e., encoded, recognized, and identified) by human listeners. For the most part, research over the last thirty-five years has demonstrated that when human observers are presented with speech signals, they typically respond to them as linguistic entities rather than simply as random auditory events in their environment. The set of labels used in responding to speech are intimately connected with the function of speech as a signaling system in spoken language. Thus, speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background and experience. And, a listener's performance in identifying and discriminating a particular acoustic attribute is often a consequence of the functional role this property plays in the listener's linguistic system. It is possible to get human listeners to respond to the auditory properties of speech signals with some training and the use of sensitive psychophysical procedures. But one of the fundamental differences between speech and nonspeech signals lies in the linguistic significance of the patterns to the listener and the context into which these patterns may be incorporated.

In the sections below, we discuss several recent findings that deal with the dynamic or time-varying aspects of speech perception. The topics to be considered in this paper include findings on trading relations, perceptual integration of acoustic cues, and context effects. The findings from these studies point to significant gaps in our current understanding of the perception of speech and nonspeech sounds in isolation and in context. At the present time, we do not have a psychophysics of speech nor do we have a psychophysics of complex sounds. Current theoretical efforts represent only a very meager beginning and, in some cases, an unsatisfactory attempt to understand a wide variety of phenomena in the field of speech perception. Our discussion of these selected topics in speech perception is designed to emphasize the wide separation that currently exists between researchers working in the mainstream of speech perception and those attempting to develop a psychophysics of speech and other complex sounds. It is hoped that this presentation will generate a great deal of discussion at the workshop about future directions for research and theory in speech perception and the research goals of investigators who are currently interested in the psychophysics of both speech and nonspeech signals.

Cue Trading and Acoustic Cue Integration. It has been well-known for many years that several cues may signal a single phonetic contrast (e.g., Delattre, Liberman, Cooper, and Gerstman, 1952; Denes, 1955; see Repp, 1982, for a review). Thus, it is possible to demonstrate that when the perceptual utility of one cue is attenuated, another cue may take on primary effectiveness in signaling the contrast under scrutiny because both cues, it is assumed, are *equivalent*. This is called a phonetic trading relation (Repp, 1982). In recent years, phonetic trading relations have been cited as evidence for a specialized speech mode of perception. There appear to be two reasons for this view. First, some demonstrations of phonetic trading relations involve both spectral and temporal cues that are distributed over a relatively long temporal interval. Repp (1982) has argued that it is hard to imagine how such disparate cues arranged across relatively long time windows could be integrated into a unitary percept if specialized (i.e., non-auditory) processes were not in operation. Repp proposes, furthermore, that the basis of this specialization lies in the listener's abstract knowledge of articulation. In other words, because we as listeners know (implicitly) how speech is produced, we are able in some way to integrate acoustically different cues that arise from an articulatory plan into a single unified phonetic percept. The second line of evidence for specialization of speech perception involves demonstrations that phonetic trading relations do not apparently arise for nonspeech sounds. Such evidence is therefore taken to be proof that the integration of multiple cues giving rise to trading relations is somehow or another peculiar to processing speech signals.

One frequently investigated trading relation involves the so-called stop manner contrast in word pairs such as "say"-"stay" or "slit"-"split." The presence or absence of a stop in such minimal pairs may be signalled by one of two cues: (1) silent closure duration between the offset of /s/ frication and onset of voicing and (2) the first formant transition onset. Fitch, Halwes, Erickson, and Liberman (1981) examined the degree to which these two cues are phonetically equivalent in perception. A demonstration of the phonetic equivalence of these two diverse cues would suggest the operation of specialized processes that "ignore" the acoustic diversity of these cues and integrate them into a unitary phonetic percept.

Fitch et al.(1981) synthesized two syllables, one having formant transitions biasing perception of the syllable /lIt/and another the syllable /plIt/. /s/ frication was appended to the beginnings of each syllable and two series of stimuli were generated by varying the closure interval between the /s/ frication and the vocalic portion of each syllable. One series of stimuli was thus composed of /s/ + /lIt/ and another of /s/ + /plIt/, with both series varying in the duration of the closure interval. Fitch et al. presented these sets of stimuli to subjects for identification. For both series, stimuli with sufficiently short closure durations were heard as /slIt/ and stimuli with sufficiently long closure durations were heard as /splIt/. Thus, Fitch et al. demonstrated that in spite of the formant transitions, the duration of the closure interval could induce indentification of the stimuli from both series as either /slIt/ or /splIt/. However, their results also showed that, on the average, relatively more silence (approximately 20 msec) was required for identification of /splIt/ for the /s/ + /lIt/ series than for the /s/ + /plIt/ series. These findings demonstrate that formant transition cues and closure duration trade off in producing perception of the presence or absence of the stop /p/.

To determine more precisely if formant transitions and closure duration are perceptually equivalent, Fitch et al. carried out a second experiment on the discrimination of /slIt-splIt/ stimuli containing either only one cue, two cooperating cues, or two conflicting cues. The logic behind this experiment was as follows: If formant transitions and closure duration are equivalent, their perceptual effects should be additive. Thus, relative to a baseline condition, adding a cooperating cue should enhance discriminability, whereas adding a conflicting cue should decrease discriminability due to the fact that the perceptual effect of these two cues should cancel one another out. This result is precisely what Fitch et al. found. Discrimination was best for the cooperating cue stimuli, intermediate for the single cue stimuli, and worst for the conflicting cue stimuli.

To further buttress the claim that phonetic trading relations (and the concommitant notion of phonetic equivalence) are peculiar to speech processing, Best, Morrongiello, and Robson (1981) performed an experiment using sine-wave analogs of "say" and "stay," a contrast for which they demonstrated a similar trading relation to that of "slit"-"split." [Sine-wave analogs were constructed by imitating the center frequencies of formants of natural speech tokens with pure tones (Remez, Rubin, Pisoni, and Carrell, 1981).] Two versions of stimuli were constructed: In one, the sine-wave portion of the stimulus had a low onset of the lowest tone (simulating /deI/ or, in Best et al.'s terms, "strong" [deI]) and one a high onset of the lowest tone (simulating /eI/ or "weak" [deI]). Noise was then appended to the beginning of each stimulus to simulate /s/ frication and test continua were generated by varying the closure interval.

Best et al.(1981) presented these stimuli to subjects for identification using an AXB procedure. In this procedure, A and B are endpoints of the continuum and X any one of the items from the continua. Subjects respond by indicating whether X is more like A or B. According to post-hoc interviews of the subjects, the subjects were partitioned into two groups, "speech" listeners and "nonspeech" listeners. [For sine-wave stimuli modelled after natural speech, some listeners spontaneously hear the stimuli as speech (although somewhat unnatural speech). Other listeners, however, hear the stimuli as nonspeech whistles (see Remez et al., 1981).] Identification functions for the "speech" or "say"-"stay" listeners revealed a trading relation; those who failed to hear the stimuli as speech, however, failed to display identification functions indicative of a trading relation. In addition, the subjects who heard the stimuli as nonspeech were further subdivided into two groups, one group which attended to spectral cues (i.e., onset frequency of the lowest tone) and one which attended to temporal cues (duration of the closure interval). Thus, the nonspeech listeners were unable to trade the two cues and attended to either the spectral cue or the temporal cue. Apparently, subjects who heard the stimuli as speech perceived the stimuli in a phonetic mode in which the temporal and spectral cues were somehow integrated into a unitary percept, thus giving rise to the observation of a trading relation; those subjects hearing the stimuli as nonspeech were presumably perceiving the stimuli in an auditory mode in which integration of the two cues was impossible.

The demonstration of trading relations constitutes the newest source of evidence for the existence of a specialized speech mode in which knowledge of articulation comes to bear in the perception of speech. According to Repp (1982), "trading relations may occur because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic information by perceiving the most plausible articulatory act. This explanation requires that the listener have at least a general model of human vocal tracts and of their ways of action" (p. 95). Thus, based in part on demonstrations of phonetic trading relations, researchers, particularly those associated with Haskins Laboratories, have once again renewed their efforts to argue for articulation-based specialized phonetic processing. It is not clear, however, that such a position is entirely unassailable.

Massaro and his colleagues (Massaro and Oden, 1980; Oden and Massaro, 1978; Massaro and Cohen, 1977) offer an alternative account of trading relations that explicitly denies any specialized processing. Instead, in their model, speech perception is viewed as a "prototypical instance of pattern recognition" (Massaro and Oden, 1980, p. 131). Briefly, Massaro and Oden argue that multiple features corresponding to a given phonetic contrast are extracted independently from the waveform and then combined in the decision processor according to logical integration rules. These rules operate on fuzzy sets so that information regarding a given feature may be more-or-less present or "sort of" present. This aspect of their model, then, stresses continuous rather than all-or-none information. Thus, features are assigned a probability value between .0 and 1.0 indicating the extent to which a given feature is present. Subsequently, the degree to which this featural information matches a stored prototype is determined according to a multiplicative combination of the independent features. The fact that multiple features are evaluated independently, and that these features can assume ambiguous values (e.g., .5), can account for the finding that the perceptual utility of two cues may trade off in rendering a given phonetic percept.

Although Massaro's model can handle phonetic trading relations without reference to articulation or specialized phonetic processing, the results of Best et al. (1981) demonstrating cue trading for speech stimuli but not for sine-wave analogs of speech present a problem. If speech perception is simply a prototypical example of pattern recognition, why are some patterns (e.g., speech) processed differently than other patterns (e,g., sine-wave analogs)? One additional, reasonable assumption can be invoked to account for the speech-nonspeech findings, namely that experience with speech stimuli sensitizes the listener to the existence of many possible redundant cues to a given phonetic contrast. For speech stimuli, then, the listener is biased toward evaluation and integration of all possible cues. For sine-wave analogs, with which the listener has presumably had little experience, the listener may hold no expectations of the possible dependencies among cues.

Thus, the absence of expectations of cue dependencies for nonspeech stimuli may have produced the differences in cue-trading effects of speech and nonspeech observed in the Best et al. study. Indeed, the very fact that some listeners in this study attended to spectral characteristics of the stimuli and others to temporal characteristics attests to the fact that both cues were available to the nonspeech listeners. However, because these subjects did not treat these stimuli as speech-like, they may not have applied certain overlearned strategies for evaluating and integrating diverse cues to the sine-wave analogs (see Grunke and Pisoni, 1982; Schwab, 1981).

Repp, Liberman, Eccardt, and Pesetsky (1978) dismiss a similar explanation of cue trading and integration on a priori grounds. However, Repp (1983) has recently argued that cue trading results do not in fact support the claim that speech is processed by specialized mechanisms. Much in the spirit of our discussion of the cue-trading literature, Repp concludes that cue-trading effects "are not special because, once the prototyptical patterns are known in any perceptual domain, trading relations among the stimulus dimensions follow as the inevitable product of a general pattern matching operation. Thus, speech perception is the application of general perceptual principles to very special patterns" (p. 132).

In short, we believe, as was shown several years ago with categorical perception effects, that reasonable alternative explanations are possible for the cue-trading evidence reported thus far. Whether these explanations will stand the test of time is, of course, an empirical question. The arguments proposed for the existence and operation of specialized speech processing mechanisms and the mediation of articulation in speech perception have quite broad implications for linguistics, psychology, and the philosophy of mind. Thus, because of the important ramifications of these claims, it is probably best to err on the side of caution in evaluating the available evidence. The cue-trading evidence is one of the most compelling sources of evidence to date for a speech mode of perception. However, the historical lesson taught by the failures of previous lines of research (e.g., categorical perception, selective adaptation) to demonstrate specialized speech processing emphasize the importance of maintaining a healthy skepticism in evaluating any new evidence for specialization of speech processing with reference to articulatory mediation. In short, we are sympathetic to the position being advocated here but we are not yet convinced from the experimental data used to support these claims.

Context Effects in Speech Perception. Much, if not all, of the research on speech perception over the last thirty-five years has been concerned with the minimal cues, features, or acoustic attributes that support perception of segmental phonetic contrasts in highly restricted environments (e.g., CV syllables). Although this reductionism has made scientific investigation more tractable, it has led many researchers to ignore, or at least postpone consideration of the perceptual problems posed by the production of speech in fluently-articulated sentences or passages of connected discourse. At the level of fluent continuous speech, the problems of invariance and segmentation appear to become even more imposing. Not only are segments coarticulated within syllables in continuous speech, but coarticulatory effects are spread across words, making isolation of words within sentences a seemingly insurmountable task for the listener. In addition, many suprasegmental effects found in continuous speech introduce other sources of variability that need to be accounted for in the perceptual process. For example, phrasal and sentential contexts introduce variations in fundamental frequency, stress placement and timing, and duration, all of which fall under the rubric of "prosodic phenomena." Finally, context effects produced by differences in speaking rate and speaker characteristics (e.g., sex, age, dialect) introduce further sources of variability that affect the acoustic-phonetic encoding of the linguistic message in the speech waveform.

Traditionally, context-conditioned variability has been viewed as a source of "noise" in the acoustic signal from which phonetic segments are extracted. Recently, however, research efforts have focused on discovering the systematic effects of context, giving rise to the notion of "lawful variability" (Elman and McClelland, 1983). Basically, this conception of contextually-conditioned variability treats context effects as sources of important acoustic-phonetic information rather than simply noise in the signal (Church, 1983; Elman and McClelland, 1983; Nakatani and O'Connor-Dukes, 1980). The notion of "lawful variability" stems from a number of diverse demonstrations of the orderliness and predictability of context effects in the production and perception of speech. What was once thought to be noise that must be filtered out in recovering phonetic segments from the waveform is now coming to be thought of as a source of useful information arising from systematic, rule-governed contextual effects.

Inherent in the idea of "lawful variability" is the growing tendency to view the speech waveform as a rich source of acoustic-phonetic information. Previously, it was thought that acoustic-phonetic information was so impoverished that higher-levels of knowledge must continually be brought to bear in the perception of speech. Recent approaches that take advantage of rule-governed variability, however, emphasize the richness and informativeness of the acoustic-phonetic information in the waveform. Thus, a number of researchers have begun to advocate more bottom-up approaches to the speech perception process, an understandable turn of affairs in light of the claim that "rather than a bane, phonetic variability may be a boon in speech perception" (Nakatani and O'Connor-Dukes, 1980, p. 13).

In the sections below, we discuss four types of context effects that have been of recent interest: (1) local phonetic context, (2) phonological and lexical context, (3) phrasal and sentential context, and (4) speaking rate. Each of these areas has proven to be highly amenable to experimental investigation and has advanced our knowledge considerably about systematic context effects in speech perception.

Local Phonetic Context Effects. One of the most pervasive effects of local phonetic context is that of allophonic variation. Allophonic variation refers to the fact that a given phoneme may have many different acoustic-phonetic realizations, depending on the context in which it is produced. For example, a /t/ in syllable-initial position, such as in /tEd/, is aspirated in English (i.e., accompanied by a short burst of noise associated with release). However, a /t/ occuring in syllable-final position is rarely released (e.g., in /bEt/) and a /t/ occurring in the cluster /st-/ is never aspirated (e.g., in /stap/). All three phonetic realizations of [t] are said to be allophones of the phoneme /t/. Although [t]'s occurring in clusters and in syllable-final position have acoustic attributes different from [t]'s occurring in syllable-initial position, we nevertheless perceive every phonetic realization of a [t] as the phoneme /t/.

A number of years ago, Nakatani and his colleagues (Nakatani and Dukes, 1977; Nakatani and Schaffer, 1978; Nakatani and O'Connor-Dukes, 1980) and Church (1983) proposed that allophonic variation should be viewed as a source of information in parsing words and syllables in sentences (see also Oshika, et al., 1975). Consider the following phonetic transcription of the question "Did you hit it to Tom?", discussed first by Klatt (1977):

[dIǰəhIɾItɨtam]

This transcription is meant to represent a "normal" articulation of the question in fluent casual speech. A spectrogram of this utterance is shown in Figure 1 along with spectrographic representations of the same words produced in isolation. As is apparent from this example, the "ideal" (or citation) forms of the words, [dId], [yu], [hIt], [It], [tu], [tam], undergo many phonetic changes when produced in sentential context. These changes, if viewed simply as noise imposed on the canonical phonetic transcriptions of the words, would appear to make the lexical retrieval process quite difficult for the listener. In particular, where does one word end and another begin? An analogous situation in printed text would arise if the spaces were removed from a sentence, such as "CATSATEASEARERARELYEARNESTOPPONENTS" ("Cats at ease are rarely earnest opponents").

Insert Figure 1 about here

The problem of parsing "Did you hit it to Tom?" into its constituent lexical items may be overcome, however, by appealling to at least two sources of information: allophonic variation and phonological contraints (Church, 1983). Church, building on the earlier work by Klatt (1977), points out that five allophonic rules are operative in the example [diǰəhIɾItɨtam]: (1) /d/ before /y/ in "did you" palatizes, rendering /dIǰu/; (2) unstressed /u/ reduces to /ə/ in "you", rendering /dIǰə/; (3) intervocalic /t/ in "hit it" flaps, rendering /hIɾ It/; (4) /u/ in "to" reduces and devoices, rendering /tɨ/; and (5) /t/ in "it to" geminates, rendering /Itɨ/. Thus, many of the phonetic changes observed in sentential context are highly predictable, and thus highly informative, if one assumes that the listener has access to implicit knowledge concerning the way allophonic variations operate. Applying these five allophonic rules, we can recover a great deal of information about the underlying phonemic representation of the sentence:

[dIǰəhIɾItɨtam] becomes /dIdyuhItIttutam/

Another source of constraint pointed out by Church is imposed by the operation of phonological rules. If we expand, as Church suggests, the original transcription to include the presence of aspiration and glottal stops, and subsequently apply a few general phonological rules, hypothesization of lexical items is further simplified. Including aspiration and glottalization renders the following transcription:

[dIǰəhIɾIʔtʰItʰam]

Syllable boundaries are now predictable given the following four phonological rules: (1) /h/ always occurs in syllable-initial position, (2) [ɾ] always occurs in syllable-final position, (3) [ʔ] always occurs in syllable-final position, and (4) [tʰ] always occurs in syllable-initial position (Church, 1983). (As Church points out, [tʰ] may be found in syllable final position, although aspiration in syllable-final position is very different from that in syllable-initial position.)

Applying the rules governing allophonic variation to recover the underlying phonemes and the phonological rules to identify syllable boundaries, we obtain the following transcription (syllable boundaries are

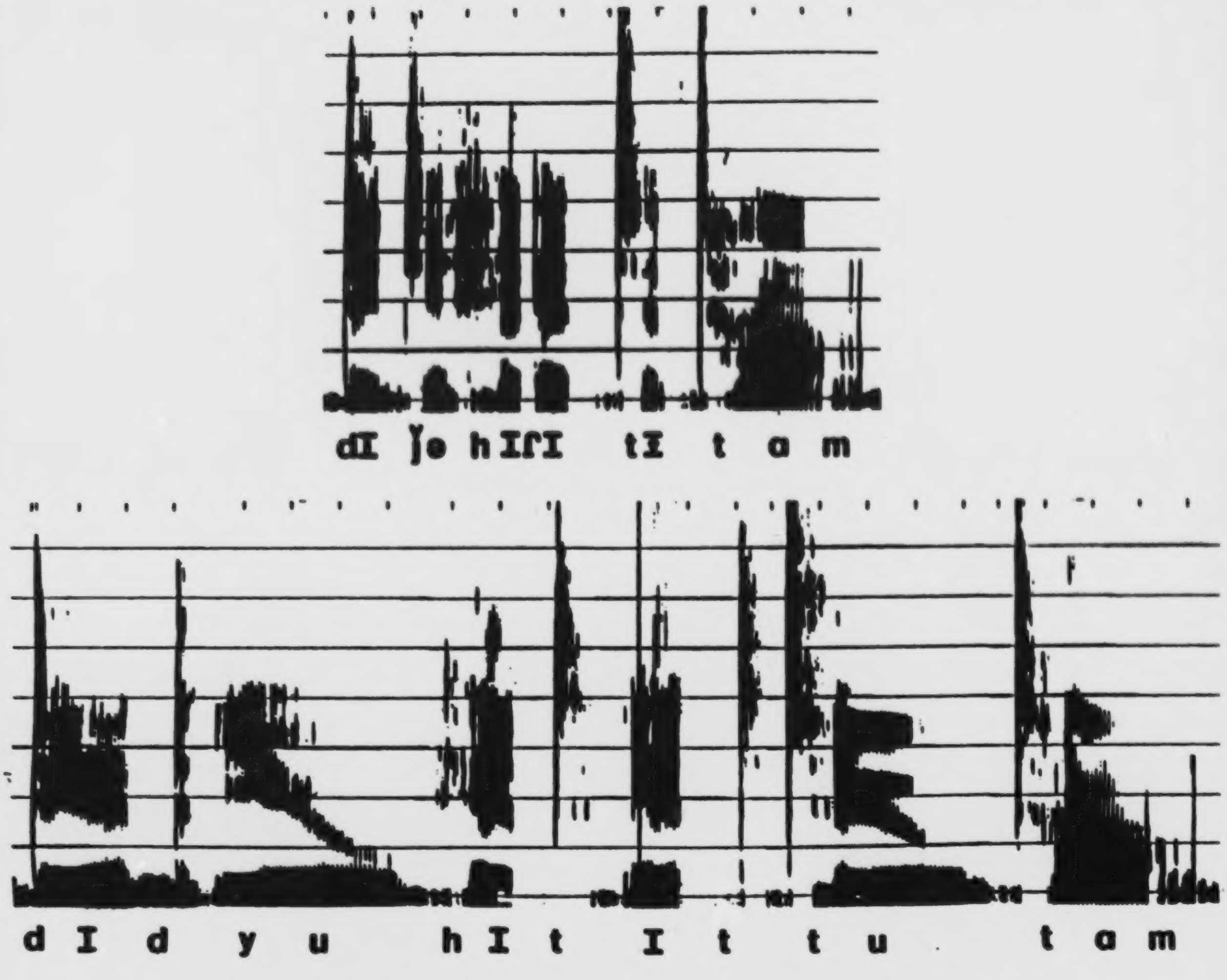


Figure 1. Spectrogram of "Did you hit it to Tom?" The same words produced in isolation are displayed in the lower panel.

indicated by a #):

/dIdyu#hIt#It#tu#tam/

It is clear from this one example that exploiting the information in a relatively fine-grained phonetic transcription allows recovery of a great deal of information concerning syllable boundaries and the underlying phonemic representations. By inference, it seems reasonable to suppose that the listener makes use of his implicit knowledge of allophonic variation and phonological rules to parse continuous speech into words. What is perhaps most compelling, however, is the degree to which the acoustic cues to allophones and syllable boundaries are differentially encoded in the signal. By treating the manifestations of allophonic variation and phonological processes in the speech waveform as sources of important information in the signal rather than simply noise, we are able to take advantage of the systematic variability contained in the speech waveform. Moreover, we are able to focus more closely on acoustic-phonetic information in resolving ambiguities, rather than having to appeal to higher-level knowledge sources (such as syntax and semantics) for "hypotheses" about what may or may not be present in the signal (Chomsky and Halle, 1968). In short, such an approach to speech perception emphasizes the richness of the speech signal instead of touting the impoverished and highly variable nature of the acoustic cues to phonetic segments and word boundaries.

One of the first researchers to advocate the notion that allophonic variation aids in parsing lexical items in sentences was Nakatani (Nakatani and Dukes, 1977). Nakatani and his colleagues have conducted a number of important perceptual and acoustic studies aimed at identifying cues to word juncture in order to specify how allophonic variation as well as prosodic information are used in identifying the beginnings and ends of words. Nakatani and Dukes (1977) examined possible allophonic cues to word juncture in pairs of words such as "no notion" and "known ocean." Such word pairs are phonetically identical except for the locus of the word juncture (see also Bolinger and Gerstman, 1957). Nakatani and Dukes excised portions of word pairs that immediately preceded and followed the word juncture and cross-spliced these excised portions between words in a pair. They then presented the spliced and original versions of these word pairs to subjects for identification. In this way, Nakatani and Dukes were able to determine whether the offset of the first word, the onset of the second word, or both contributed to perception of a word boundary.

Nakatani and Dukes' results showed that word junctures were almost entirely cued by the onset of the second word in the pair, except for words ending in /r/ or /l/. Because /r/ and /l/ have distinctly different allophones at the beginnings and endings of words, these allophones constituted strong cues for word juncture both word-initially and word-finally. In addition, Nakatani and Dukes found that allophonic variations at the beginnings of the second word in the pairs provided cues to word juncture even in the absence of /r/ or /l/. In particular, they found that glottalization and/or laryngealization cued word junctures when the second word began with a vowel. Finally, they showed that aspiration of voiceless stops, which is most evident for word initial allophones, aided in identifying word junctures.

The findings of Nakatani and Dukes (1977) provide strong empirical support for Church's (1983) claim that allophonic variation is an important source of information in segmenting words in sentences. In another study of word juncture cues, Nakatani and O'Connor-Dukes (1980) extended their previous experiment to include a number of other allophonic and segmental differences that cue word juncture. They found: (1) that gemination of consonants helped to distinguish such pairs as "drunk converse" (in which a doubling or gemination of the /k/'s is present) and "drunken verse" (in which no gemination occurs); (2) that flapped apical stops distinguish pairs such as "hardy feat" (which contains a flapped /d/) and "hard defeat" (which contains the geminate /d+d/); (3) that the presence of a syllabic /n/ distinguishes pairs such as "maiden forced" (which contains a syllabic /n/) and "maid enforced"; (4) that deletion of the unstressed vowels in words such as "bakery," pronounced "bakry," distinguishes word pairs such as "bakery guarded" and "bake regarded"; and finally (5) that vowel reduction in prefixes such as "de-" distinguishes word pairs such as "hard defeat" and "hardy feat." In short, Nakatani and O'Connor-Dukes have shown that listeners take much allophonic variation into account in parsing word strings into their constituent lexical items.

Nakatani and Schaffer (1978) and Nakatani and O'Connor-Dukes (1980) have also shown that stress patterns and rhythm can also aid in identifying word boundaries (see also Nakatani, O'Connor, and Astor, 1981). Using reiterant speech, which preserves prosodic information but eliminates allophonic variation and other segmental differences (see Liberman and Streeter, 1978), these researchers demonstrated that listeners could correctly parse reiterant speech versions of adjective-noun phrases such as "malformed nose" and "long stampede." Taken together, these studies demonstrate that much information resides in the speech signal that can significantly affect segmentation of sentences into words. What was considered by some to be noise and random variation (e.g., allophonic variation) appears to have quite important ramifications for the identification of words in fluent speech from the bottom-up analysis of the speech waveform.

Phonological and Lexical Context Effects. We have seen how local phonetic context may serve to guide a listener's parsing of sentences into words. We now turn to the issue of how somewhat higher-level linguistic constraints can influence a listener's perception of phonetic segments. In the preceding section, we suggested that knowledge of phonological rules may aid the listener in recovering underlying phonemic representations and in identifying syllable boundaries. In this section we turn to a somewhat more abstract role of phonology in speech perception, namely the role of knowledge of phonologically permissible sequences in speech perception (sometimes called phonotactics). In addition, we examine some evidence that relates to the effects of lexicality on the perception of phonemes. Both phonological and lexical context effects illustrate the degree to which the listener's knowledge of permissible sound sequences and words in the lexicon influence his or her perception of phonetic segments (see also Pisoni, Luce, and Nusbaum, 1986).

Massaro and Cohen (1983) have recently reported the results of an experiment aimed at evaluating the degree to which phonological context can affect listeners' perception of phonemes. In one of their conditions, Massaro and Cohen generated a synthetic continuum ranging from /ri/ to /li/. The manipulation of crucial interest was the consonant preceding the /ri-li/ syllables. Massaro and Cohen placed each of the /ri-li/ stimuli after one of four consonants: /p/, /t/, /s/, and /v/. In English, both /ri/ and /li/ are permissible after /p/; only /ri/ is permissible after /t/; only /li/ is permissible after /s/; and neither /ri/ nor /li/ are permissible after /v/. Massaro and Cohen were interested, in part, in determining if phonological context (permissible and non-permissible) would affect subjects labelling of the /ri-li/ continua. In particular, they hypothesized that more /r/

responses would be obtained for the /tri-tli/ continuum and more /l/ responses for the /sri-sli/ continuum.

As predicted, Massaro and Cohen found that phonological context did affect listeners' labelling of the stimuli. Their subjects produced more /r/ responses than /l/ responses in the context of /t/ and more /l/ responses than /r/ responses in the context of /s/. The identification functions for the /pri-pli/ and /vri-vli/ continua fell between the two other functions, as expected. Massaro and Cohen furthermore showed that their effect was in fact due to phonological context and not to auditory interactions between the initial stops and the following /ri/ or /li/ syllables.

In a similar experiment, Ganong (1980) examined the effect of lexical context on the identification of word-initial stops. Ganong varied the VOT of word-initial stops to generate continua ranging from a word to a nonword (e.g., "dash" to "tash") and from a nonword to a word (e.g., "dask" to "task"). He then presented these stimuli to subjects for identification. Ganong found that lexicality (i.e., whether the stimulus was perceived as a word or a nonword) strongly affected subjects' labeling of the word initial stop. Subjects produced more "dash" responses for the "dash"-"tash" continuum and more "task" responses for the "dask"-"task" continuum.

The Massaro and Cohen (1983) and Ganong (1980) studies both demonstrate the effects linguistic knowledge can have on the categorization of speech sounds. These results show that perception of phonetic segments is heavily influenced by what listeners know about permissible sequences of speech sounds in English and by their knowledge of words in the lexicon. Thus, phonological and lexical context further serve to constrain the perceptual analysis of the speech signal. These studies, in conjunction with those by Nakatani and his colleagues, also demonstrate that relatively early in the perceptual processing of speech, many ambiguities may be resolved by employment of allophonic rules, phonological rules, and lexical constraints. We view these studies as important new demonstrations of the influences of phonological and lexical context on the perception of phonetic segments. This work furthermore represents a sharp departure from the earlier views that assumed the speech perception process was strongly driven by top-down knowledge of syntax and semantics.

Sentence-level context effects. Thus far, we have discussed the systematic variability of phonetic segments in various local phonetic environments and how the listener's knowledge of allophonic variation, phonological rules, and lexical items may serve to support perception of phonemes and words. Another source of systematic variation is introduced, however, when our attention is focused beyond the phonetic segment or word to the study of speech produced in sentential contexts. At this level of analysis, sentence-level effects come into play. We use the term "sentence-level context effects" to refer to those changes in the acoustic-phonetic structure of speech that arise not from the effects of the articulation of adjacent segments, but from the production of fluent speech in sentences.

One of the most widely studied effects of sentence-level context concerns the changes in fundamental frequency (F0), duration, and amplitude that occur at phrase boundaries. The presence of a major syntactic boundary (e.g., the boundary between an initial subordinate clause and a main clause) may be signaled by any of a number of possible cues: a marked fall in in the slope of F0 preceding that boundary (Cooper and Sorenson, 1981; Maeda, 1976; Pierrehumbert, 1979), a "resetting" of F0 following the boundary (Cooper and

Sorenson, 1981; Maeda, 1976), a pronounced lengthening of segments immediately preceding a boundary (Klatt, 1975; Oller, 1973; Luce and Charles-Luce, 1985), a decrease in amplitude at the boundary (Streeter, 1978), and a pause at the boundary (Goldman-Eisler, 1972). The question addressed by many of the studies investigating syntactic boundary phenomena concerns which of these cues are most important for the listener in identifying phrase boundaries.

Much of the work on the perception of phrase boundaries has employed ambiguous utterances that are manipulated in such a way as to allow assessment of a single potential cue (e.g., Lehiste, 1973; Lehiste, Olive, and Streeter, 1976; Lehiste, 1983). The most comprehensive of these studies was performed by Streeter (1978). Streeter examined the relative importance of phrase-final lengthening, F0 declination, and changes in amplitude for the identification of phrase boundaries in ambiguous algebraic expressions. In one condition, Streeter electronically manipulated duration, F0, and amplitude of utterances of the phrase [A plus E times O], which may be read as [(A plus E) times O] or [A plus (E times O)], and required subjects to identify which of the two possible readings was intended for a given stimulus. She found that both duration and F0 served to cue phrase boundaries, whereas amplitude had little effect. (See Luce and Charles-Luce (1983) for similar findings obtained from a reaction time task.) Moreover, she found that duration and F0 were additive, not interactive, cues. Streeter's study thus demonstrates that changes in F0 and duration induced by the presence of a phrase boundary are important independent cues for listeners in the identification of phrase boundaries.

Cutler (Cutler and Darwin, 1981; Cutler and Foss, 1977) has examined the extent to which prosodic information enables the listener to predict where sentence stress will fall. Because sentence stress is usually placed on words of primary semantic importance in a sentence, the ability to predict sentence stress would presumably enable the listener to focus in on those words in a sentence most crucial to the message. Thus, prosodic variations may help direct the listener to high information centers in fluent speech.

It is clear that listeners rely on variability introduced by suprasegmental context effects to extract syntactic and semantic information from the speech waveform. Thus, duration and pitch changes caused by the occurrence of syntactic boundaries and by sentence stress placement provide valuable information for the parsing and comprehension of sentence-length utterances. Moreover, it is clear that a listener's processing of prosodic information is quite complex, in that no single cue has yet to be shown to be necessary in identifying phrase boundaries or stress placement (see Cutler and Ladd, 1983). Although the recent interest in the role of prosody in speech perception is certainly a welcome trend, much work is needed to specify more precisely how the listener takes advantage of this obviously important source of information in the perception of fluent speech.

Effects of Speaking Rate. One final effect of suprasegmental context that deserves discussion is that induced by changes in speaking rate. Effects of speaking rate on the perception of speech are not, in the strict sense of the term, "prosodic" effects. Instead, the issue of the effects of speaking rate on the perception of speech relates to the issue of "perceptual normalization." Just as we may ask how a listener compensates or normalizes for the acoustic consequences of changes in the vocal tract sizes of different speakers, we may also ask how the listener normalizes for changes in speaking rate (within and between speakers). In short, how do listeners normalize for speech produced in the context of many different speaking rates?

In a comprehensive review of the effects of "global" speaking rate on the production and perception of phonetic segments, Miller (1981) discusses a number of changes at the phonetic level induced by changes in speaking rate. For vowels, both spectral and durational changes may be observed as speaking rate is increased. In particular, vowels tend to reduce at faster rates of speech so that target formant frequencies are rarely achieved. For consonants, cues to voicing of syllable-initial (VOT) and intervocalic stops (closure duration) undergo systematic changes as speaking rate is speeded or slowed. In addition, manner class distinctions between consonants are likewise affected by changes in speaking rate.

One of the most interesting findings concerning the effects of speaking rate on the perception of segmental contrasts concerns voicing of stop consonants in syllable-initial and intervocalic position. In a series of studies, Summerfield (1974, 1975a, 1975b; Summerfield and Haggard, 1972) examined the effects of speaking rate on the identification of stimulus continua varying along the dimension of VOT. He found that the rate of articulation of the carrier sentence in which the stimuli were embedded affected the voicing boundaries for the continua in systematic ways. In particular, for a carrier sentence produced at a fast speaking rate, shorter VOT's were required to identify a stimulus as voiceless than when the carrier sentence was produced at a slower speaking rate.

On the basis of these and other studies, it appears that listeners adjust their judgments of phonetic contrasts to compensate for perceived speaking rate. Moreover, the adjustments are highly systematic and predictable. Although it is not the case that all of the effects of speaking rate heretofore demonstrated are so straightforward as those demonstrated by Summerfield (see also Port and Dalby, 1982), it is probably true that the variability introduced by changes in speaking rate are automatically compensated for by listeners (Miller, Green, and Schermer, 1982) and have highly predictable effects on listeners' perceptions. Unfortunately, we do not have a good theoretical account of these findings yet nor a deep understanding of the perceptual mechanisms responsible for this form of perceptual compensation (see, however, Pisoni, Carrell and Gans, 1983).

Conclusions. Despite these recent findings and their immediate impact on theoretical efforts in speech perception, there are still very large gaps in our understanding of the auditory/perceptual processing of speech signals by human listeners. In the past, it has been very easy to account for a set findings in speech perception by appealing to the existence and operation of specialized speech processing mechanisms. Unfortunately, such global explanatory accounts are becoming more and more unsatisfactory as we begin to learn more about the psychophysical and perceptual properties of speech and complex nonspeech signals and how the auditory system encodes these types of signals. It is clear to us that theoretical accounts of specific phenomena in speech perception such as trading relations, cue integration, and context effects can no longer be couched in terms of vague descriptions of articulatory mediation by specialized perceptual mechanisms. We have not carried out all the appropriate nonspeech control experiments yet but we are certain that more precise and testable explanations of these findings will be forthcoming in the years ahead.

References

Best, C. T., Morrongiello, B., and Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 1981, 29, 191-211.

Bolinger, D. and Gerstman, L. J. Disjuncture as a cue to constructs. Word, 1957, 13, 246-255.

Chomsky, N., and Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Church, K. W. Phrase-structure parsing: A method for taking advantage of allophonic constraints. Bloomington, Ind.: Indiana University Linguistics Club, 1983.

Cooper, W. E., and Sorensen, J. M. Fundamental Frequency in Sentence Production. New York: Springer-Verlag, 1981.

Cutler, A., and Darwin, C. J. Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. Perception and Psychophysics, 1981, 29, 217-224.

Cutler, A., and Foss, D. J. On the role of sentence stress in sentence processing. Language and Speech, 1977, 20, 1-10.

Cutler, A., and Ladd, D. R. Prosody: Models and Measurments. New York: Springer-Verlag, 1983.

Delattre, P. C., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. An experimental study of the acoustic determinants of vowel color: Observations of one- and two-formant vowels synthesized from spectrographic patterns. Word, 1952, 8, 195-210.

Denes, P. Effect of duration on the perception of voicing. Journal of the Acoustical Society of America, 1955, 27, 761-764.

Elman, J. L., and McClelland, J. L. Exploiting lawful variability in the speech waveform. Paper presented at the Symposium on Invariance and Variability, M.I.T., Cambridge, Mass., 1983.

Fitch, H. L, Halwes, T., Erickson, D. M., and Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception and Psychophysics, 1980, 27, 343-350.

Ganong, W. F. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 110-125.

Goldman-Eisler, F. Pauses, clauses, sentences. Language and Speech, 1972, 15, 103-113.

Grunke, M. E., and Pisoni, D. B. Some experiments on perceptual learning of mirror-image acoustic patterns. Perception and Psychophysics, 1982, 31, 210-218.

Klatt, D. H. Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 1975, 3, 129-140.

Klatt, D. H. Review of the ARPA speech understanding project. Journal of the Acoustical Society of America, 1977, 62, 1345-1366.

Lehiste, I. Phonetic disambiguation of syntactic ambiguity. Glossa, 1973, 7, 107-121.

Lehiste, I., Olive, J. P., and Streeter, L. A. The role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America, 1976, 60, 1199-1202.

Liberman, M. Y., and Streeter, L. A. Use of nonsense-syllable mimicry in the study of prosodic phenomena. Journal of the Acoustical Society of America, 1978, 63, 231-233.

Luce, P. A., and Charles-Luce, J. The role of fundamental frequency and duration in the perception of clause boundaries: Evidence from a speeded verification task. Journal of the Acoustical Society of America, 1983, 73, S67.

Luce, P. A., and Charles-Luce, J. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. Journal of the Acoustical Society of America, 1985, 78, 1949-1957.

Maeda, S. A characterization of American English intonation. Unpublished doctoral thesis. Cambridge, Mass.: M.I.T., 1976.

Massaro, D. W., and Cohen, M. M. The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America, 1977, 60, 704-717.

Massaro, D. W., and Cohen, M. M. Phonological context in speech perception. Perception & Psychophysics, 1983, 34, 338-349.

Massaro, D. W., and Oden, G. C. Speech perception: A framework for research and theory. In N. J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice, Vol. 3. New York: Academic Press, 1980, 129-165.

Miller, J. L. Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.), Perspectives on the Study of Speech. Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

Miller, J. L., Green, K., and Schermer, T. On the distinction between prosodic and semantic factors in word identification. Journal of the Acoustical Society of America, 1982, 71, UU6.

Nakatani, L. H., and Dukes, K. D. Locus of segmental cues for word juncture. Journal of the Acoustical Society of America, 1977, 62, 714-719.

Nakatani, L. H., and O'Connor-Dukes, K. D. Phonetic parsing cues for word perception. Unpublished manuscript. Murray Hill, NJ: Bell Laboratories, 1980.

Nakatani, L. H., and Schaffer, J. A. Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, 1978, *63*, 234-245.

Nakatani, L. H., O'Connor, K. D., and Aston, C. H. Prosodic aspects of American English speech rhythm. *Phonetica*, 1981, *38*, 84-106.

Oden, G. C., and Massaro, D. W. Integration of featural information in speech perception. *Psychological Review*, 1978, *85*, 172-191.

Oller, D. K. The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, 1973, *54*, 1235-1247.

Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H., and Aurbach, J. The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975, *ASSP-23*, 104-112.

Pierrehumbert, J. The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, 1979, *66*, 363-369.

Pisoni, D. B., Carrell, T. D., and Gans, S. J. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, 1983, *34*, 314-322.

Pisoni, D. B., Luce, P. A., and Nusbaum, H. C. The role of the lexicon in speech perception, this volume.

Pollack, I. The information of elementary auditory displays. *Journal of the Acoustical Society of America*, 1952, *24*, 745-749.

Port, R. F., and Dalby, J. Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics*, 1982, *32*, 141-152.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. C. Speech perception without traditional speech cues. *Science*, 1981, *212*, 947-950.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 1982, *92*, 81-110.

Repp, B. H. Trading relations among acoustic cues in speech perception: Speech-specific but not special. *Haskins Laboratories Status Report on Speech Research* SR-76. New Haven: *Haskins Laboratories, 1983, 129-132.*

Repp, B. H., Liberman, A. M., Eccardt, T., and Psestsky, D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *1978*, *4*, *621-637*.

Schwab, E. C. Auditory and phonetic processing for tone analogs of speech. Unpublished doctoral dissertation. Buffalo, NY: State University of New York at Buffalo, 1981.

Stevens, K. N. Acoustic correlates of some phonetic categories. Journal of the Acoustical Society of America, 1980, 68, 836-842.

Streeter, L. A. Acoustic determinants of phrase boundary perception. Journal of the Acoustical Society of America, 1978, 64, 1582-1592.

Summerfield, Q. Towards a detailed model for the perception of voicing contrasts. In Speech Perception (No.3). Belfast: Department of Psychology, Queen's University of Belfast, 1974.

Summerfield, Q. Cues, contexts, and complications in the perception of voicing contrasts. In Speech Perception (No. 4). Belfast: Department of Psychology, Queen's University of Belfast, 1975. (a)

Summerfield, Q. Information processing analysis of perceptual adjustments to source and context variables in speech. Unpublished doctoral dissertation. Belfast: Department of Psychology, Queen's University of Belfast, 1975. (b)

Summerfield, Q., and Haggard, M. P. Speech rate effects in the perception of voicing. In Speech Synthesis and Perception (No. 6). Cambridge: Psychology Laboratory, University of Cambridge, 1972.

Using Template Pattern Structure Information to Improve Speech Recognition Performance*

Moshe Yuchtman and Howard C. Nusbaum

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*This research was supported by NIH Research Grant NS-12179 to Indiana University. An earlier version of this paper appeared in the *Proceedings of the American Voice Input/Output Society*, September 1986. We thank Chris Davis for his assistance in carrying out some of this research.

Abstract

Isolated utterance, speech recognition systems are typically trained by collecting speech data for each of the words in some vocabulary for a particular application. The speech data is then converted into a template or model of the pattern structure of each word. During recognition, each input utterance is converted into the same type of pattern representation and this representation is compared to all the patterns that were stored during training. The distance or similarity of the input pattern to the trained patterns is determined and a recognition decision is based on applying criteria for minimally acceptable similarity scores. However, the formation of templates during training and the recognition decision are not based on all the information available in the vocabulary. We found that patterns of similarity among utterances in the vocabulary can be obtained during training, and these similarity patterns, in turn, may accurately predict the structure and probabilities of confusions during recognition. In this paper, we describe a method for incorporating intra-vocabulary similarity into an improved decison-making process.

Using Template Pattern Structure Information to Improve Speech Recognition Performance

The vast majority of commercially available speech recognition systems must be trained on the speech of a particular talker prior to recognizing that talker's utterances. To train a recognition system, the talker produces one or more tokens of each vocabulary item, and the recognizer constructs a template or model of each word in the vocabulary based on these speech samples. These stored representations of the vocabulary serve as the reference patterns against which utterances are compared for recognition. During recognition, speech input is converted into the same type of pattern representation and the similarity or distance of this input pattern is determined by comparing it to all the stored representations. The recognition decision is based on applying criteria for minimally acceptable similarity or distance scores to each of the vocabulary items. The recognized word is the item with the smallest distance or the greatest similarity score. Thus, there are several general components to the operation of most recognition systems: (1) a training or enrollment procedure that is used to construct representations of each of the vocabulary items, (2) a pattern matching procedure that compares an input utterance with the stored representations to compute similarity scores, and (3) a decision procedure that selects the best candidate based on the similarity scores.

These very general principles of operation are typically quite effective for successful recognition of vocabularies composed of acoustically distinctive utterances. Thus, vocabularies consisting of words varying in length, stress location, phonotactic pattern, and phonetic constituency will be recognized with relatively low error rates. By comparison, much higher error rates may result for recognition of acoustically similar items such as the E-set (i.e., B,D,G,P,T,C,Z, and E). Poor performance on phonetically similar items is largely due to the poor phonetic resolution of current speech recognition systems. Improvements in phonetic recognition algorithms will undoubtedly improve overall recognition performance for phonetically confusable vocabularies. However, it is also possible that the performance of the current generation of recognition algorithms may be improved by taking into account information about the structural properties of the vocabulary (see Pisoni, Nusbaum, Luce, and Slowiaczek, 1985).

Almost all recognition systems treat the words in a vocabulary as if they were completely independent of each other. The current generation of isolated-utterance, speaker-dependent speech recognition systems ignore the structural relationships among words in a vocabulary. During training, the only information used in constructing a representation of a word is the separate tokens of the word that were produced by the talker. In deciding which word was spoken, speech recognition systems use one or two simple criteria such as the distance score for the highest candidate and the difference in distances between the two highest candidates. But there is a great deal more information available in acoustic-phonetic and lexical structure of the vocabulary that could be used to improve recognition performance. For example, after a vocabulary is enrolled (i.e., the recognizer has been trained on one token of each vocabulary item), the similarity of the templates to each other could be used to modify the training procedure to increase the distinctiveness of the representation of similar words. In addition, in trying to decide which of two similar candidates should be chosen as the correct recognition response, a decision algorithm could examine the pattern of distance scores to all items in the vocabulary. The pattern of distance scores distributed throughout the vocabulary could be quite informative about the correct candidate.

To make effective use of the pattern of distance scores among the items in a vocabulary, it is necessary to quantify this information with some metric. Multidimensional scaling provides an analysis technique that describes the structure of distances and confusions within a fixed vocabulary. In general, a multidimensional scaling analysis yields a geometric configuration which, for a given dimensionality, best describes (in terms of variance accounted for) a set of observed distances among stimuli. In our research, we have used a scaling procedure based on the INDSCAL model (Carroll and Chang, 1970) that allows for a simultaneous analysis of several distance matrices. An n-dimensional solution provides two sets of data: (1) the projections of stimuli on each dimension which is referred to as the "stimulus space", and (2) the relative weights of each dimension for each individual matrix which is called the "subjects" or "condition space." In the domain of speech recognition both types of information are of considerable importance.

By applying multidimensional scaling techniques to the analysis of confusions and intra-vocabulary distances, it is possible to determine the acoustic properties of a vocabulary that are most and least distinctive for a particular speech recognition algorithm. This provides a metric of the "recognition space" for an algorithm. An examination of this recognition space may provide diagnostic information about the performance of a speech recognition system, since this type of analysis indicates precisely along which dimensions some subsets of a vocabulary are confused and along which dimensions other subsets are discriminated. By examining projections of stimuli onto these dimensions it is be possible to determine which acoustic properties correspond to these dimensions. Thus, based on a multidimensional scaling analysis of one vocabulary, it may be possible to predict speech recognition performance for a different vocabulary that is constructed along similar dimensions.

Structure in the Recognition Space

In a recent study (Yuchtman, Nusbaum, and Davis, 1986), we carried out a multidimensional scaling analysis of the recognition errors and distances between words in the E-set vocabulary for several commercially available speech recognition systems. We found that four-dimensional Euclidean spaces provide an excellent fit to observed inter-word distances. These four-dimensional spaces account for approximately 80 to 88 percent of the variance in recognition performance. An example of this type of stimulus configuration is shown in Figure 1. This figure displays the projections of the E-set on the first and second dimensions of the 4-dimensional solution obtained for one speech recognition system. Because the items in the E-set vocabulary differ primarily in the initial consonant segment, it is possible to interpret these dimensions in straightforward acoustic-phonetic terms.

Insert Figure 1 about here

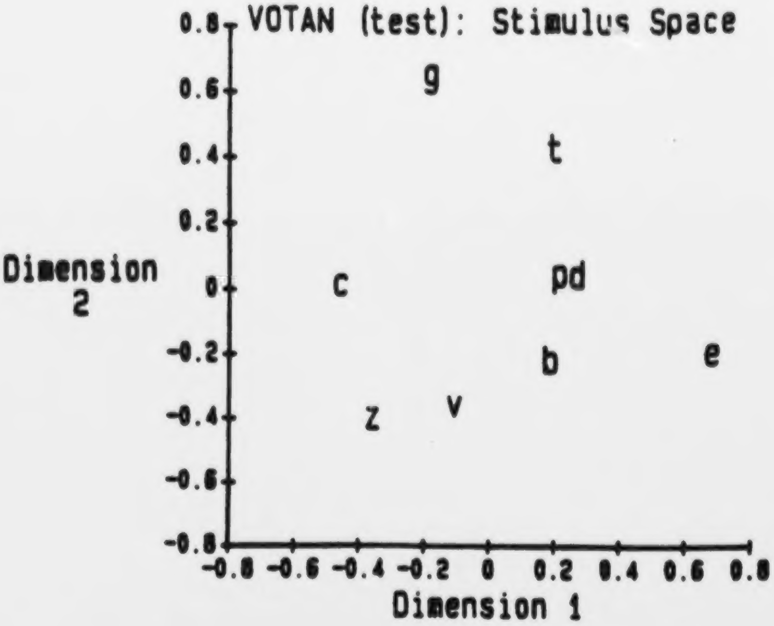


Figure 1. Word projections on dimension 1 and 2 of the SINDSCAL solution for inter-word distances measured during recognition (VOTAN recognizer).

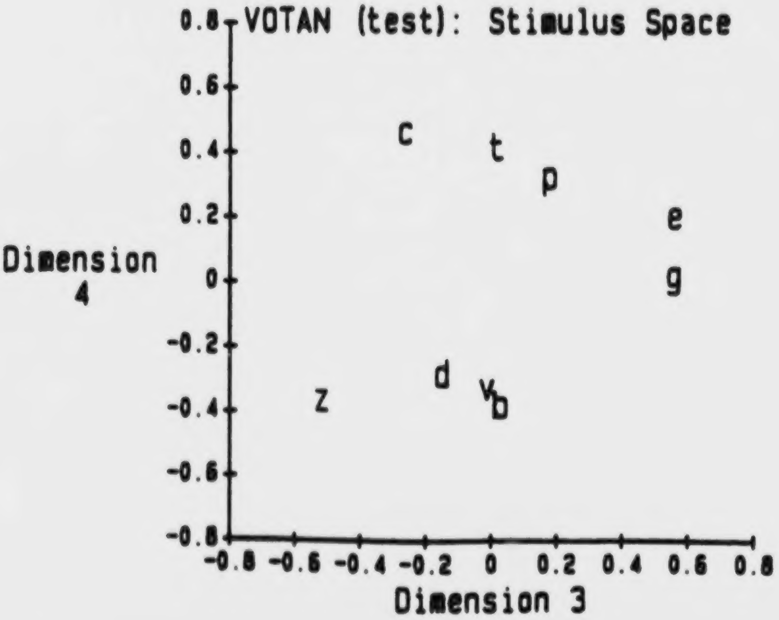


Figure 2. Word Projections on dimension 3 and 4 of the SINDSCAL solution for inter-word distances measured during recognition. (VOTAN recognizer).

The first dimension assigns the stimuli to four broad categories: (1) C,Z; (2) G,V; (3) the stops P,B,T,D; and (4) the vowel E. Clearly, the projections along this dimension correspond to the level and/or duration of frication noise in the word. On the second dimension, the stimuli G and T are placed at one extreme, while Z, V, B, and E are located on the other. This dimension may be interpreted in terms of the contrast between sounds with abrupt onset of noise vs. sounds with gradual onset of periodicity or noise.

The projections of stimuli on the third and fourth dimensions are shown in Figure 2. The distribution of words along the third dimension produces a sequence in which, by moving from negative to positive weights, we find the stimuli C and Z, followed by their homorganic stops, T and D. The labial sounds P,B, and V assume a location in the center, while E and G have the highest positive weights. In general, this sequence corresponds to differences in spectral peaks and formant transitions that are associated with the place of articulation of the initial sound segments.

Insert Figure 2 about here

The fourth dimension classifies stimuli according to the voicing value of the initial consonants: B,V,D, and Z vs. C,T, and P, with G and E assuming a middle position. The projection of E on this dimension, places this vowel in the neighborhood of the voiceless consonants. An acoustically based account for this observation is not immediately apparent, although it may reflect the presence of a glottal stop preceding the vowel.

In general, the scaling solutions obtained for two other speech recognition systems can also be interpreted in terms of some underlying acoustic properties that differentiate between several subsets of the E-set (e.g., fricatives and non-fricatives). It is important to note, however, that the fine details of the scaling solutions vary among different recognizers. As one might expect, different speech recognition algorithms will use different coding schemes and pattern representations so the structure of confusions will differ. In the scaling solution, this will be revealed in differences in the specific dimensions and dimensional weights that account for these confusions. In this way, the multidimensional scaling analysis of recognition data may directly reflect the idiosyncratic properties of signal processing algorithms of different recognition systems. Thus, this approach provides important diagnostic information about the way in which recognition spaces may differ. This approach suggests one method of predicting performance for application vocabularies based on performance data from tests carried out with a laboratory "benchmark" vocabulary.

It may be possible to construct a special "calibrated vocabulary" (see Ohala, 1982) that contains words differing systematically along various structural dimensions (e.g., length, stress pattern, phonotactic pattern, phonetic constituency). Multidimensional scaling analyses of recognition performance for this type of calibrated vocabulary could provide important information about the sensitivity of a particular recognition algorithm to each of the dimensions that characterize the calibrated vocabulary. Once the salient dimensions are known for a specific recognition system, it should be possible to predict performance on any vocabulary by analyzing the new vocabulary to determine the degree to which these salient dimensions are represented in the new vocabulary. For example, if word length is an important dimension for a recognition algorithm and phonetic constituency is not, performance should be better for vocabularies consisting of words differing in length compared to vocabularies consisting of words differing in the identity of specific phonemes.

Predicting Patterns of Similarities from Training Data

Beyond the capability of analyzing the structure of recognition spaces and predicting performance for new vocabularies, it is also possible to use multidimensional scaling to predict the pattern of confusions for recognition of a vocabulary based solely on the training data. In the present study, we compared multidimensional scaling solutions obtained for similarity scores derived from training tokens and similiarity scores derived from test tokens.

Method

The recognition vocabulary used in this study consisted of the E-set of the alphabet. The actual speech tokens are part of the Texas Instruments database collected by Doddington and Schalk (1981). This database was produced by eight male and eight female talkers with 10 training tokens and 16 test tokens per word. Three commercially available speech recognition systems were tested: (1) the VOTAN VPC-2000, (2) the Interstate Vocalink CSRB, and (3) the NEC SR-100.

To determine the distances among all the vocabulary items for each test, a recognition device was trained using a single vocabulary item and then tested on the entire vocabulary. In other words, for one test, a recognizer was trained on tokens of B and then was tested on all letters in the E-set. Using this paradigm, we determined directly the distances among all the vocabulary items.

Distances between each template in the vocabulary, and each training or test token for the E-set, were obtained directly from the recognizer. Triangular distance matrices were constructed from the mean values of these distances for the training (n=5), and test tokens (n=16). Separate scaling analyses were conducted for each recognizer and distance type using the SINDSCAL program (Pruzansky, 1975). In all cases, the input consisted of 16 matrices constructed for each talker, and solutions were obained in 1-to-6 dimensional configurations.

Results

Figure 3 shows the first two dimensions of the scaling solution generated for the VOTAN. In this figure, the lower-case characters represent data for recognition of the training tokens, while the upper-case characters represent projections obtained from recognition of the test tokens. Training and test data for the third and fourth dimensions are shown together in Figure 4.

Insert Figure 3 and Figure 4 about here

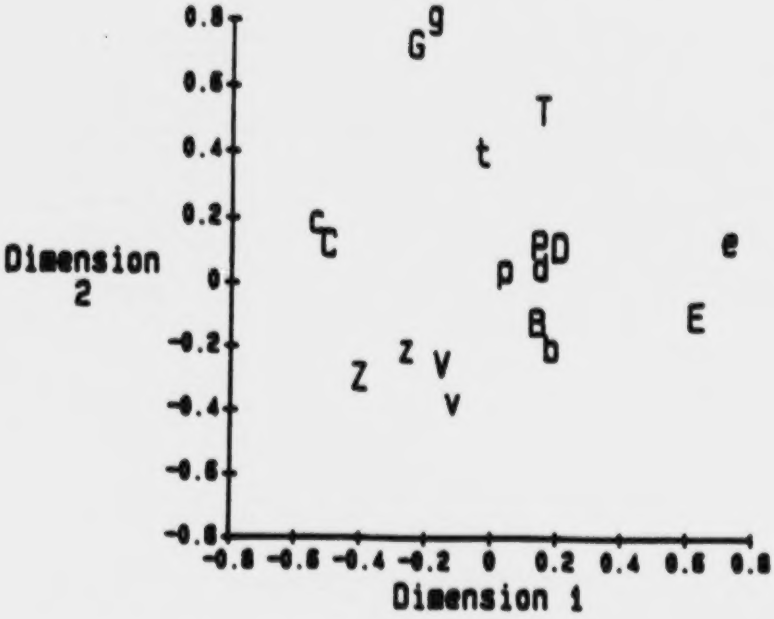


Figure 3. Word Projections on dimension 1 and 2 of SINDSCAL solutions for distances measured during training (lower-case) and recogniton (upper-case) for the VOTAN recognizer.

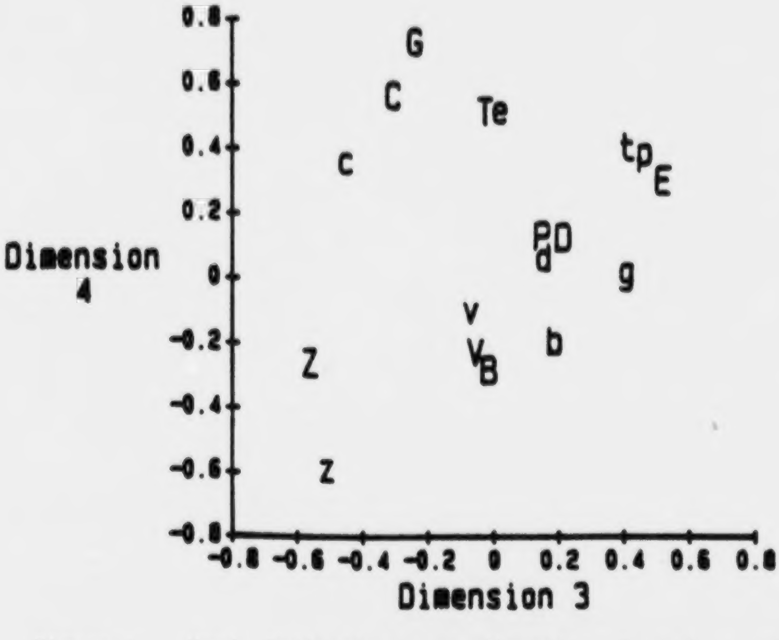


Figure 4. Word Projections on dimension 3 and 4 of SINDSCAL solutions for distances measured during training (lower-case) and recogniton (upper-case) for the VOTAN recognizer.

Clearly, the patterns of stimulus projections on both dimensions are remarkably similar for the training tokens and the test tokens. In several cases, the order or polarity of a particular dimension had to be changed in order to achieve maximum overlap. However, it is important to note that distances among stimuli are invariant over these transformations. Excellent agreement between training- and test-based solutions was observed for the other two recognizers as well. The value of the product-moment correlation between the test and training projections obtained for all three devices, and for each of the four dimensions. is .92, indicating that the similarity patterns among the training tokens predicts about 81% of the variance in the similairty patterns of the test tokens.

This consistency in the scaling solutions for independent sets of data (i.e., the training tokens and the test tokens) has two important implications. First, this finding indicates that scaling solutions obtained for the three recognizers are highly reliable. Thus, differences between recognition spaces may be safely attributed to the operational characteristics of each device rather than measurement error. Second, it is evident that the distances beween vocabulary items used in the process of training can accurately predict the pattern of distances between the same items during actual recognition trials. It should be noted that distances among stimuli in the spaces yielded by multidimensional scaling are normalized and relative rather than absolute. While actual error probabilities can be inferred from such data indirectly, they may be derived using simpler computational procedures. The importance of the data yielded by multidimensional scaling, and possibly by other multivariate approaches is in the insight scaling analyses provide about the interaction between the operational characteristics of a recognition system and the acoustic attributes of its application vocabulary as produced by individual talkers. Moreover, in the case of highly confusable items, this information may suggest which acoustic property can be used to optimize discriminability.

Predicting Recognition Accuracy from Training Data

In the previous section we have shown that patterns of similarities (i.e., the confusions) among words in the E-set obtained from recognition data can be reliably predicted from inter-word distances measured during training. Using a set of training and recogniton data obtained in a "simulated" recognition experiment, we examined the extent to which intra-vocabulary distances obtained during training can reliably predict recogniton accuracy for indivdual talkers or words. The recognition experiment was simulated in the sense that no speech recognition device was used in this study. Instead, a recognition system was simulated using software subroutines on a VAX-11/750 computer.

Method

The recognition vocabulary used in this study consisted of the E-set from the TI database produced by eight male talkers. Five tokens of each word were used for training. All 16 test tokens were used for recognition testing.

Simulated training and recognition was carried out using ILS Version 5.0 (Signal Technology Inc., 1985) software. Linear predictive cepstral coefficients were computed every 12.8 msec and were used for subsequent training and recognition. Template generation, as well as distance measurements, were carried out using a non-linear warping alogorithm. Following training, distances between all training tokens and templates were measured and stored for later use. Likewise, distances between each test token and all templates were computed and stored for subsequent analysis.

Results

Overall, tokens of the E-set were recognized with an accuracy of 76.4%. This performance falls well within the range of performance obtained using the commercially available speech recognition systems for this database. Thus, the performance of our simulated speech recognition system is comparable to that obtained for actual recognition devices.

As in the previous study, the distances among words measured during training are highly correlated with distances obtained during recognition test trials. This indicates that the distances among tokens presented for training is highly predictive of the pattern of confusions that will result during subsequent recognition testing.

Another parameter that predicts the pattern of confusions during recognition is the ratio of the distances between the first and second word candidates for an utterance. Almost all speech recognition systems will return more than one possible candidate as a recognition response and, in many cases, the correct word is contained within the first two candidates. When the first candidate is very similar to the input utterance and the second candidate is very different, recognition accuracy will be very high. Conversely, when the two candidates have similar distance scores, accuracy will be very poor. Figure 5 shows, for each of the 8 talkers, mean values of proportion of correct responses as a function of the mean distance ratios measured during training. The correlation bewteen the two variables is relatively low (.72), primarily because one talker deviated considerably from the rest of the talkers.

Insert Figure 5 about here

In Figure 6, we plotted, for each of the 9 words in the E-set, mean percent correct word recognition obtained during recognition test, as a function of the ratio of the distances for the first and second candidates measured during training. The high correlation between the two series (.92), suggests that recognition accuracy for individual words can be reliably predicted using measures of similarity beween training tokens.

Insert Figure 6 about here

Taken together, the results indicate that it is not only possible to predict the pattern of confusions from distance scores obtained during training, but it is also possible to predict the actual level of recognition accuracy. If it is possible to predict, prior to recognition, which items will be confused and how often they will be confused, it should also be possible to use this information to reduce the error rate and improve overall recognition performance. From a detailed analysis of the recognition space

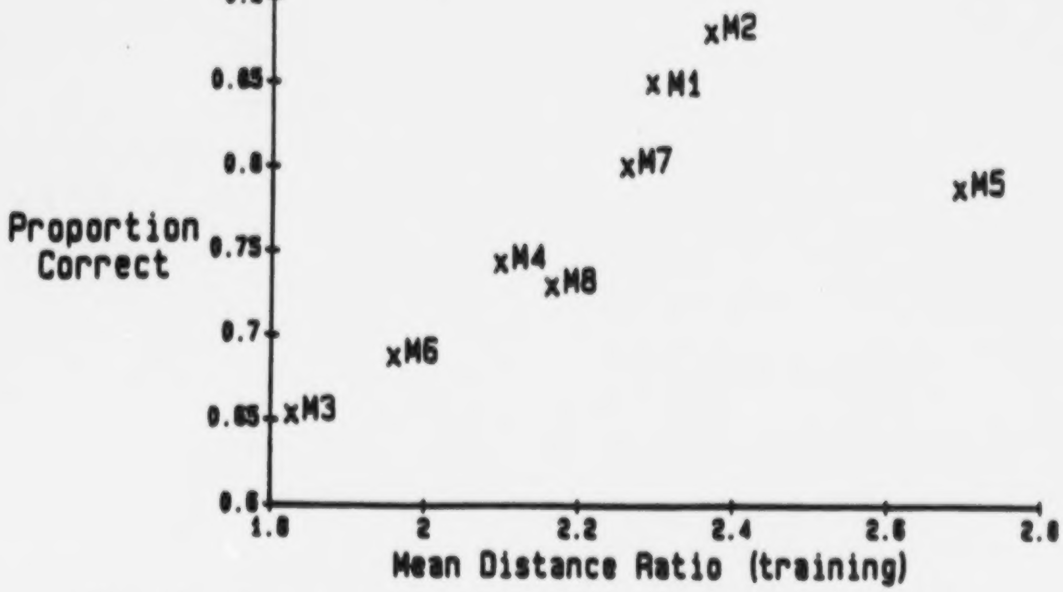


Figure 5. Mean proportion of correct decisions for individual talkers as a function of distance ratios measured during training.

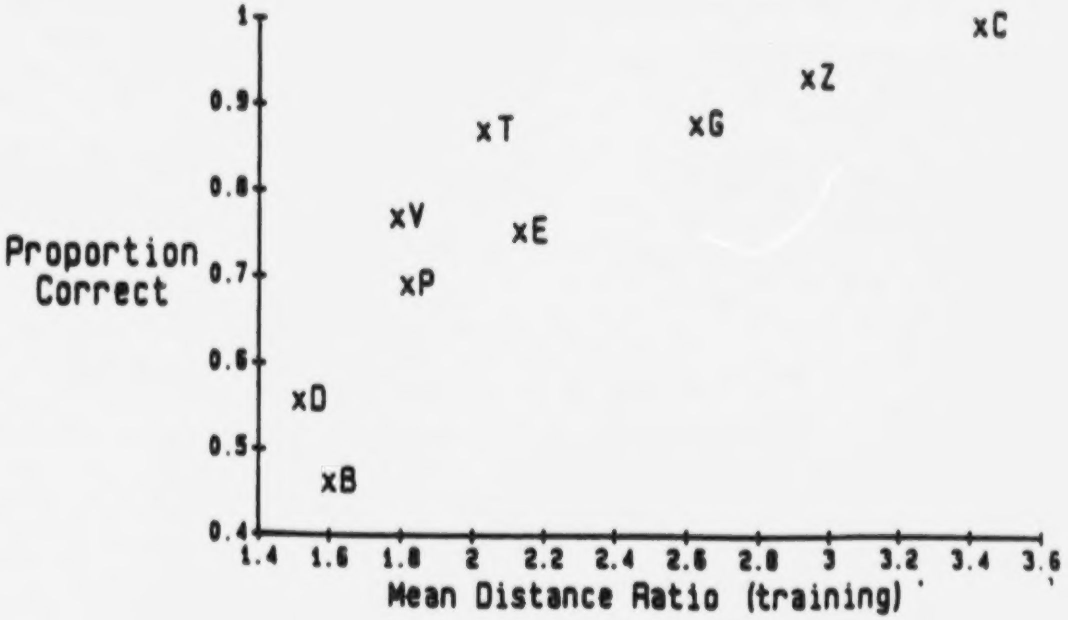


Figure 6. Mean proportion of correct decisions for each word in the E-set as a function of distance ratios measured during training.

during a first pass of training, we can determine which vocabulary items will be confused. This type of analysis may also suggest how to refine the acoustic analysis of utterances processed during subsequent training passes to produce templates or word models that are more distinctive. This type of modification of the training protocol based on the structure of the recognition space is just one way of improving recognition performance. Another method is to use information that predicts the structure of confusions to enhance the decision algorithm that selects one candidate as a recognition response out of the two or more choices returned by the recognizer. Currently, most recognition systems will return the first candidate if its similarity or distance crosses some threshold (i.e., the reject threshold). Raising the reject threshold trades substitution errors (incorrect responses) for rejection errors (no recognition response returned). Another decision strategy is to require the difference between the distance scores for the first and second candidates to exceed some criterion (i.e., the delta reject threshold). This strategy works for the reason outlined above namely, that the ratio of first and second candidate distance scores is an accurate predictor of performance. However, few if any currently available speech recognition systems provide a decision strategy that can choose the second candidate over the first. The reject threshold and delta reject strategies only serve to filter out the first candidate based on a comparison of its distance score with a standard reference or with the second candidate. In the next study we examined the operation of a decision strategy designed to choose between the first and second candidates.

Incorporating Word Neighborhood Structure into Decision Making Rules

From the data presented so far, it is quite clear that the distances between words, as measured using a small set of training tokens, provides a good estimate of both the frequency and types of confusions that occur during actual recognition. As discussed earlier, this information may be of considerable "diagnostic" value by indicating which subsets of a given vocabulary may be highly confusable and by predicting patterns of confusions based on calibrated vocabularies. In addition, the pattern of distances within a fixed vocabulary may be used to improve the performance of a recognition system. In particular, when two candidates are both similar to an input utterance, the pattern of distance scores for other vocabulary items may be used to aid in deciding which candidate is the correct response.

Data from the simulated recognition of the E-set vocabulary, decribed above, were used in the present investigation. The training data included a set of inter-template distances and mean distances between training tokens and templates for each talker and for each word. Recognition data for each talker and word consisted of a set of the distance values measured between each of 16 test tokens and the nine stored templates.

Decision Rules

We examined the performance of three decision rules on recognition performance for our simulation. The first of these was the minimal distance rule by which an utterance is identified as being a token of the most similar template (i.e., the template with the smallest distance from the utterance). In some recognition systems, the minimal distance decision is made if the distance does not exceed pre-established criteria (i.e., reject and delta reject thresholds).

In addition to the minimal distance rule, we examined two other rules that incorporate different quantitative properties of similarity structures (as measured during training) into the decision making process. Both of these rules rely on the distances obtained during training among training tokens and the distances between the input utterance and all the vocabulary templates. Rule 1 selected either the first or second word candidate based on which candidate produced a larger product-moment correlation score computed for the nine distances between the utterance and each of the vocabulary items and the series of distances obtained for the training tokens used in training each candidate. According to this decision rule, a candidate is selected as the correct recognition response, based on the candidate that yields a higher correlation between the training data and the utterance-to-template distance scores for the entire vocabulary. Rule 2 is similar to Rule 1 in that it is based on correlations between the training distances and the utterance-to-template distances. The difference between Rule 1 and Rule 2 is that Rule 2 is based on a rank-order (Spearman) correlation, instead of the product-moment correlation.

Results

The performance of each of the three decision rules (minimum distance and Rules 1 and 2) was evaluated using several criteria for applying the rules. In all cases, the evaluation of these rules was limited to cases in which the correct recognition response was one of the first two candidates. Since the correct word had to be the first or second candidate, not all errors produced by a recognition system could be remediated by these decision rules. However, 51% of the total number of errors consisted of cases in which the correct response was one of the first two two candidates.

The performance of these rules was first examined using all the data within this constraint. The minimum distance rule alone recognized the E-set with an accuracy of 86.5%. By comparison, Rule 1 (product-moment correlation) was slightly lower in overall accuracy -- 84.0% correct, while Rule 2 (rank-order correlation) yielded a considerably lower score -- 71.8% correct. At first glance, this suggests that the minimum distance criterion used by most speech recognition systems is the optimal decision rule. However, the minimum distance rule and the two correlation rules take into account slightly different types of information. Thus, the words that are correctly recognized using the correlation rules may not be the same as those recognized by the minimum distance rule.

If the different decision rules are indeed orthogonal, it should be possible to invoke one rule when another rule is likely to produce an error and thus improve overall recognition accuracy with the composite recognition decision. In particular, if the minimum distance rule is generally quite accurate, an auxilliary correlation rule should only be invoked when the minimum distance rule will be inappropriate and error prone. There are two conditions when this is likely to happen: (1) when the utterance is about equidistant from two or more templates, and (2) when a high distance value is measured between an utterance and its nearest template. In the first case, any delta reject decision will produce a rejection error and in the second case a reject threshold rule will produce a rejection if the distance is too great. Thus, these are cases that are likely to produce rejection errors in an application. Moreover, in both these cases there is no apparent criterion for choosing another candidate since the second candidate would also fail by the reject threshold and delta reject criteria.

We tested the performance of the three decision rules using two different criteria for their application. In one evaluation, correlation rules were applied if the ratio of distances measured between an utterance and its two nearest templates was below a certain cutoff. That is, when the distance scores for two candidates were too similar, the correlation rules were invoked. The cummulative proportion of correct responses yielded by each rule for varying values of the distance ratios is shown in Figure 7. As expected, the minimum distance rule yields a higher proportion of correct responses as the distance ratio increases. At a distance ratio of 1.05, the minimum distance rule yields an accuracy of 42%. When the distance ratio is 1.5, accuracy improves to 77%.

Insert Figure 7 about here

By comparison, it can be seen in this figure that the correlation rules are less sensitive to the distance ratio between nearest candidates. Indeed, for distance ratios lower than 1.15 both correlation rules yield better results than the original minimum distance decison rule. Also, it is quite clear from this figure that the product-moment correlation rule (Rule 1) consistently outperforms the rank-order correlation rule (Rule 2). Rule 1 seems especially powerful at extremely low distance ratios: At a distance ratio of 1.05, Rule 1 yields 62.5% correct responses compared to 52% correct for Rule 2, and 42% correct for the minimum distance decision rule.

A second evaluation of these rules was carried out in which we manipulated the absolute distance between an utterance and its nearest template. Specifically, criterion cutoffs were defined as ratios between recognition distance and the mean training distance obtained for the identified word. In general, a low value of this parameter reflects a high degree of similarity between an utterance and a template and thus should be associated with a higher probability of a correct decision. The three rules were examined for recognition-to-training distance ratios ranging between 1 to 3. The results are shown in Figure 8. It is apparent that within a wide range of recognition distances, each rule performs at a constant level. Within this range, the minimum distance rule and Rule 1 are within about 1-2 percentage points of each other and both are about 10-15 percentage points better than Rule 2. However, at recognition-to-training ratios exceeding 2.0, a drop can be observed in the proportion of correct responses made using the minimum distance decision rule, while for these values, Rule 1 is more accurate by a margin of 3-4 percentage points.

Insert Figure 8 about here

Several conclusions are suggested by the results of our analyses. As expected, when an utterance is about equally distant from two templates, decisions made on the basis of a minimum distance rule are likely to result in a high error rate. This is also true, to a lesser extent, if the utterance cannot be closely matched by any template in the vocabulary. Under these

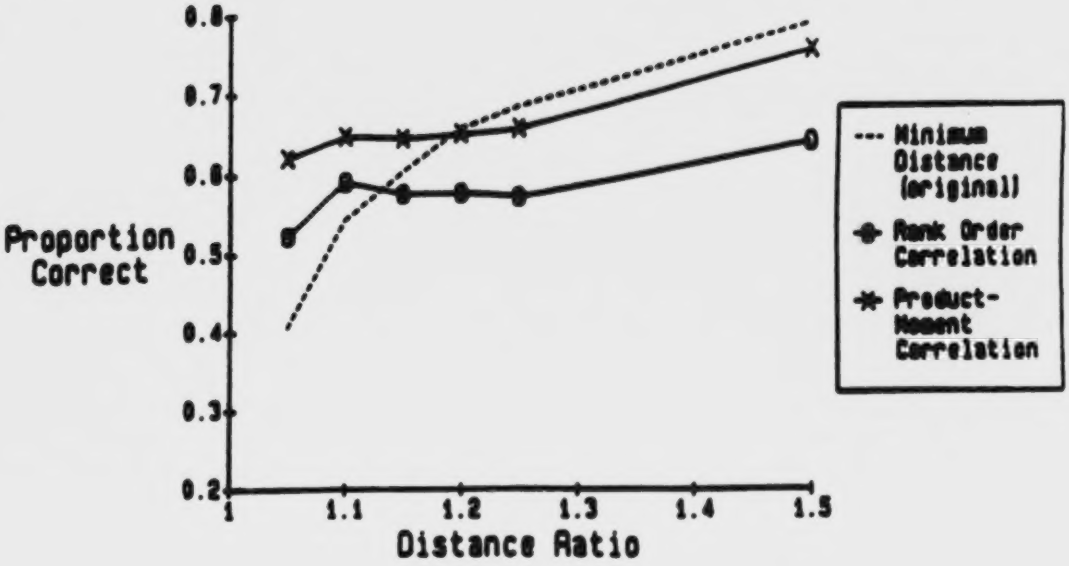


Figure 7. Mean proportion of correct decisions generated by each decision rule as a function of cutoff values of distance ratios.

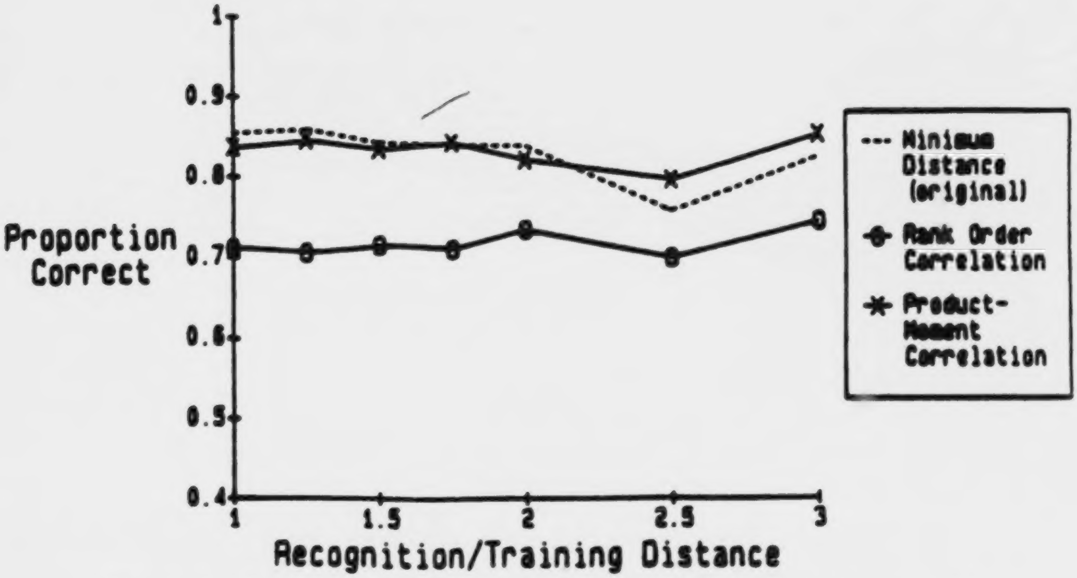


Figure 8. Mean proportion of correct decisions generated by each decision rule as a function of cutoff values of recognition-to-training distance ratio.

conditions, a rule that utilizes information about the structure of words in a similarity neighborhood as measured during training, provides a more robust decision-making mechanism. From the differences in performance between the rank-order and the product-moment correlation based rules, it is evident that lower error rates are observed if the neighborhood structure is quantified using a more detailed representation.

Overall, conditions that lead to high error rates using the minimum distance rule, will also generate a higher proportion of incorrect decisions by the neighborhood rules. However, by applying neighborhood decision rules when minimum distance decisions are least effective, recognition performance can be improved. Moreover, it is likely that further research may reveal more effective means of using lexical neighborhood data in the decision-making process. However, more accurate recognition of acoustically similar utterances may result if training information is incorporated directly into the recognition algorithm itself.

General Discussion

Almost all speech recognition systems treat the items in a vocabulary as completely independent entities. While a recognition system may take into account the pattern similarities and differences among the different tokens of a particular vocabulary item, few, if any, consider the structural properties of the vocabulary as a whole. However, without examining the phonetic similarity of the items in a vocabulary, it is difficult for a recognition algorithm to determine which portions of a word are most distinctive (since this may be determined by the context of the vocabulary). Furthermore, by not analyzing the pattern of confusions across all vocabulary items, current speech recognition systems ignore a source of information that is very relevant to selecting the correct recognition response.

Taken together, the present results argue for the importance of quantifying the structure of the recognition space produced for a specific speech recognition system and vocabulary. By using multidimensional scaling analyses of distance scores and confusion errors, it is possible to characterize this recognition space and provide useful information for predicting and improving the performance of recognition algorithms. We believe that these analysis techniques can be very useful in predicting the performance of a recognition system for new and untested vocabularies. Furthermore, these analyses can provide new methods for improving recognition performance *without modifying existing recognition algorithms*, by simply improving the decision strategies used to select the correct recognition response.

While these techniques are very powerful and offer much more promise for characterizing and improving recognition performance than we have illustrated here, we believe that the real significance of these techniques will be realized with the next generation of speech recognition systems. There are currently no established standardized methods for testing and comparing the performance of commercially available speech recognition systems. The recognition systems that are currently available have vocabulary sizes that are generally under 256 words. Within a very short time, however, several recognition systems will be available with vocabularies of 5,000 words or more. The performance of these new and more powerful systems cannot be completely described by testing with databases of digits or spoken letters. Furthermore, collecting many different 5,000 word databases using large numbers of talkers to test performance of these systems will simply not be practical. By using multidimensional scaling analyses together with a specially designed diagnostic database using a "controlled vocabulary", it may be possible to adequately describe the performance characteristics of a large vocabulary speech recognition system. Moreover, it may be possible to predict performance on any application vocabulary based on a linguistic analysis of the application vocabulary in comparison with the diagnostic vocabulary. Finally, beyond the issues of performance measurement and prediction, multidimensional scaling may suggest new ways of improving training, recognition, and decision algorithms. We have already demonstrated how the decision algorithm of a small vocabulary speech recognition simulation can be improved by developing metrics based on neighborhood similarity. Further research is needed to determine if an extension of these techniques of neighborhood analysis to large vocabulary speech recognition will provide comparable improvements in recognition performance.

References

Carroll J. D., and Chang J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of the "Eckart-Young" decomposition. Psychometrica, 283-319.

Doddington G. R., and Schalk T. (1981). Speech recognition: Turning theory into practice. IEEE Spectrum, 18, 26-32.

Interactive Laboratory System (ILS) User's Guide V5.0. (1985). Speech Technology, Inc. Goleta, Ca.

Ohala, J. (1982). Calibrated vocabularies. In D. Pallett (Ed.), Proceedings of the Workshop on Standardization for Speech I/O Technology. Gaithersberg, MD: National Bureau of Standards.

Pisoni, D. B, Nusbaum, H. C., Luce, P. A., and Slowiaczek, L. M. (1985). Speech perception, word recognition, and the structure of the lexicon. Speech Communication, 4, 75-95.

Pruzansky S. (1975) How to Use SINDSCAL: A Computer Program for Individual Differences in Multidimensional Scaling. Murray Hill, N.J: Bell Telephone Laboratories.

Yuchtman, M., Nusbaum, H. C., and Davis, C. N. (1986). Multidimensional scaling of confusions produced by speech recognition systems. Journal of the Acoustical Society of America, 79, S95-S96.

On Word-Initial Voicing: Converging Sources of Evidence in Phonologically Disordered Speech*

Judith A. Gierut

Speech Research Laboratory
Department of Psychology

Daniel A. Dinnsen

Department of Linguistics

Indiana University
Bloomington, IN 47405

---

*This research was supported, in part, by NIH Training Grant NS-07134 and NIH Grant NS20976 to Indiana University in Bloomington. We would like to thank Peter Alfonso, Jan Charles-Luce, Karen Forrest, Judith Johnston, David Pisoni, and Moshe Yuchtman for their suggestions and assistance in preparation of this manuscript. We would also like to thank Lisa Kapper for assisting with the spectrographic analyses.

Abstract

The purpose of this study is to bring related sources of data, i.e., phonological and acoustic phonetic, to bear on the characterization of two children's disordered phonological systems. Auditorily-based phonological analyses indicated that the children exhibited a superficially similar pattern of error involving the voice contrast in word-initial obstruent stops, even though both children accurately produced the voice contrast in post-vocalic stops. Acoustic phonetic analyses indicated, however, that one of the children systematically affected the voice distinction using closure duration and voice onset time, whereas the other child did not. Despite the similarity of their errors as assessed by auditorily-based phonological analyses, the children had very different productive knowledge of word-initial voicing in stops. These findings have implications for the clinical assessment and treatment of children with phonological disorders.

On Word-Initial Voicing: Converging Sources of Evidence in Phonologically Disordered Speech

Studies of both normal phonological acquisition (Barton & Macken, 1980; Kornfeld & Goehl, 1974; Macken & Barton, 1977, 1980; Menyuk, 1972) and phonological disorders (Hoffman, Stager, & Daniloff, 1983; Maxwell, 1981a, 1981b; Maxwell & Weismer, 1982; Weismer, Dinnsen, & Elbert, 1981) have indicated that young children often produce contrasts among sounds that are not perceived by adult listeners. Children mark phonological contrasts by producing phonetic distinctions, which may or may not be comparable to those used by adults. Fine-grained acoustic phonetic analyses are needed in such cases to determine whether a systematic distinction is being made. Phonetic distinctions in the absence of perceptible phonological contrasts have been taken as evidence that children have more knowledge of the sound system than is immediately apparent to listeners at an auditory level. A child's knowledge of the sound system, as used herein, refers specifically to productive knowledge or the ability to produce systematic phonetic (articulatory and/or acoustic) distinctions among sounds for the purpose of marking phonological contrasts in the language. A child's knowledge of the sound system can also be evaluated on the basis of other skills, such as speech perception (cf. Menn, 1983). These skills, however, lie outside of the domain of speech production and are possibly even independent processes (cf. Dinnsen, 1984, 1985; Straight, 1980).

For phonologically disordered children, in particular, these findings are relevant to the clinical assessment and treatment of speech sound errors. With regard to clinical assessment, these findings suggest that, in some cases, auditorily-based descriptions of a child's speech sound errors may be inaccurate (Maxwell & Weismer, 1982; Weismer, 1984; Weismer et al., 1981). Maxwell and Weismer (1982), for example, reported the case of a child who did not evidence a voice, place, or manner contrast among word-initial obstruents; phonologically, this child only produced [d] in word-initial position. Acoustic phonetic evidence, however, indicated that the child produced a three-way distinction among obstruents. This child demonstrated more productive knowledge of the sound system than was available from the phonological analysis. Conventional data for the phonological analysis were not sensitive enough to identify these subtle, systematic phonetic differences among word-initial obstruents.

With regard to clinical intervention, these findings suggest that children with similar patterns of error may have very different productive knowledge of sounds, and therefore, may require different treatments. For example, Weismer, Dinnsen, and Elbert (1981) observed three children who displayed superficially similar phonological errors involving omission of word-final obstruents. Acoustic phonetic data indicated that two of the children marked final obstruents in terms of vowel duration differences, even though final obstruents were omitted. These children evidently had productive knowledge of final obstruents,[1] but used a phonological rule of word-final deletion.[2] The third child, on the other hand, did not produce or otherwise acoustically mark final obstruents. This child exhibited a pattern of error characterized by inaccurate lexical representations of morphemes (relative to the target); that is, the child apparently represented morphemes without post-vocalic obstruents. Consequently, a rule of word-final deletion would not be applicable. Notice that these three children all displayed superficially similar errors, but their productive knowledge of sounds was different. It is expected that these differences in productive knowledge would affect the goals of treatment, such that two of the children would need to be taught to eliminate a phonological rule, while the third child would need to alter the underlying lexical representation of morphemes.

Phonological and acoustic phonetic evidence serve an important function in the accurate assessment of a phonologically disordered child's productive knowledge of the sound system and in the identification of an appropriate treatment plan. To date, however, there have been relatively few studies which have reported the use of related sources of data in the assessment or treatment of phonological disorders. With the exception of Weismer et al., there have been no other studies which have established differences in productive knowledge for children exhibiting similar patterns of error; moreover, differences in productive knowledge have not been reflected in subsequent treatment goals. Thus, the purpose of this paper is two-fold: (1) to present related sources of data (phonological and acoustic phonetic) in descriptions of the sound systems of two misarticulating children, and (2) to identify differences in productive knowledge and treatment despite superficial similarities in error pattern. Two studies follow. The first study analyzed phonologically the speech of two children and, thus, established the similarity of their error pattern. The second study was motivated by the first and, thus, analyzed acoustically the apparent phonological similarity between both children.

Subjects

Two children, Aaron, age 4 years, 6 months, and Becca, age 4 years, 3 months, participated as subjects. Both children were functional misarticulators, producing errors on several sounds from different sound classes, as determined by performance on the Goldman-Fristoe Test of Articulation (Goldman & Fristoe, 1969). The children were especially suited for this study given the similar nature of their errors in production of the voice contrast in word-initial stops. The children were from monolingual English-speaking homes, and had no previous history of language, hearing, cognitive, or motor disorders.

Phonological Analyses

Data collection

A spontaneous speech sample, 30 minutes in duration, was elicited individually from each child in varied situations, such as play and story-telling. The spontaneous speech sample was supplemented by a probe sample. The probe sample consisted of single-word spontaneous productions elicited through picture and object naming tasks. The probe sample ensured that each child had ample opportunity to produce all target English sounds; it also allowed for the elicitation of potential minimal pairs and morphophonemic alternations.

All speech samples were tape-recorded, then narrowly transcribed (IPA notation) and glossed by the first author. If utterances could not be glossed, transcriptions were still included in the data set, providing information about the occurrence and distribution of sounds.

Data analysis

Standard generative phonological descriptions were developed for each child, consistent with procedures outlined by Dinnsen (1984), Kenstowicz and Kisseberth (1979), and Maxwell and Rockman (1984). Each child's sound system was described in terms of the phonetic and phonemic inventories, distribution of sounds, lexical representation of morphemes, phonological rules, and phonotactic constraints.

Intrajudge reliability

Ten percent of each child's spontaneous speech sample used in the phonological analysis was retranscribed by the first author approximately nine months after it was obtained. Point-to-point intrajudge reliability was calculated for all consonants produced. Mean intrajudge agreement for Aaron's sample was 85% (N = 459 segments), and for Becca's sample, 94% (N = 431 segments).

Results

Aaron's phonological system. Aaron maintained a limited phonetic inventory with regard to the target sound system. His phonetic inventory included the sounds:

m	n	ŋ	
pb	td	kg	ʔ
	s		
	t͜s d͜z		
w			h
	r		

Of the fricatives, Aaron never produced [f,v,θ,ð,z,ʃ] in any word position; instead, he used [s]. In fact, [s] was the only fricative Aaron ever produced. Affricates were produced, i.e., [t͜s, d͜z]; however, they were not those of the adult sound system, i.e., [ʧ,ʤ].

Stops were produced in intervocalic and final positions, and these sounds did not alternate, as shown in the following forms:

Target [p]	[wipi]	~	[wip]	"zipping" - "zip"
	[tupi]	~	[tup]	"soupy" - "soup"
Target [b]	[wʌbɪŋ]	~	[wæp]	"rubbing" - "rub"[3]
	[kɔbi]	~	[kɔp]	"cobby" - "cob"
Target [t]	[buɾi]	~	[but]	"bootie" - "boot"
	[bætɚ]			"butter"
	[la͜ɪ t]			"light"
Target [d]	[badi]			"body"
	[pa͜ɪdɚ]			"spider"
	[a͜ɪd]			"hide"
	[da͜ɪd]			"find"

Target [k]	[dæki]	~	[dæk]	"duckie" - "duck"
	[kɪkuən]	~	[kɪk]	"kicking" - "kick"
Target [g]	[dɪgi]	~	[dɪg]	"digging" - "dig"
	[ɛgi]	~	[ɛg]	"egg-i" - "egg"

Moreover, a voice contrast[4] was maintained for intervocalic and final stops as illustrated by these minimal and near minimal pairs:

"supper"	[dʌpɚ]	-	[dzʌbu]	"jumping"
"tape"	[tep]	-	[seb]	"shave"
"water"	[ɔtɚ]	-	[odɚ]	"over"
"ice"	[a͜ɪt]	-	[a͜ɪd]	"hide"
"leggy"	[ɛki]	-	[ɛgi]	"egg-i"
"frog"	[pwɔk]	-	[pwæg]	"flag"

In word-initial position, Aaron also produced stops. However, there was no evidence of a voice contrast in that position; voiced and voiceless stops freely varied:

[p] ~ [b]	[pɪgi]	~	[bɪgi]	"piggie"
	[pinæ̃t]	~	[binæ̃t]	"peanut"
	[pwʌsən]	~	[bwʌtən]	"brushing"
[t] ~ [d]	[ti]	~	[di]	"teeth"
	[top]	~	[dop]	"soap"
	[teobo͜ʊ]	~	[deobo͜ʊ]	"sailboat"
[k] ~ [g]	[kom]	~	[gom]	"comb"
	[kɛt]	~	[gɛt]	"catch"
	[kɔt]	~	[gɔt]	"got"

The absence of a word-initial voice contrast was also evident in the child's homophonous production of potential minimal pairs:

"pie"	[ba͜ɪ]	-	[ba͜ɪ]	"bye"
"pig"	[bɪg]	-	[bɪg]	"big"
"tie"	[da͜ɪ]	-	[da͜ɪ]	"die"
"town"	[da͜ʊn]	-	[da͜ʊn]	"down"
"coat"	[ko͜ʊ]	-	[ko͜ʊ]	"goat"
"cow"	[ka͜ʊ]	-	[ka͜ʊ]	"gown"

These data indicate, among other things, that Aaron maintained a phonemic voice contrast among stops in the intervocalic and final positions of words, but not in word-initial position.

Becca's phonological system. Becca maintained a relatively complete phonetic inventory with regard to the adult system. Her phonetic inventory included the sounds:

m			n			ŋ	
pb			td			kg	ʔ
	fv	θð	sz	ʃ			
				tʃ	dʒ		
w					j		h
			l		r		

Becca produced target fricatives [f,v,θ,ð,s,z,ʃ] in post-vocalic positions. She did not, however, produce these same sounds in word-initial position; stops were used instead. The affricates patterned in a similar manner to the fricatives; [tʃ,dʒ] were used post-vocalically but not word-initially.

Like Aaron, Becca produced stops in the intervocalic and final positions of words. These stops did not alternate, as shown in the forms:

Target [p]	[tupi]	~	[tup]	"soupy" - "soup"
	[dɪpi]	~	[dɪp]	"chippy" - "chip"
Target [b]	[tʌbi]	~	[dʌb]	"tubby" - "tub"
	[wo͜ʊbi]	~	[wo͜ʊbs]	"robe-i" - "robes"
Target [t]	[bæɾi]	~	[bæt]	"fatty" - "fat" [5]
	[ba͜ɪɾɪn]	~	[ba͜ɪt]	"biting" - "bite"
Target [d]	[widɪn]	~	[wid]	"reading" - "read"
	[mʌdi]	~	[mʌd]	"muddy" - "mud"
Target [k]	[bʊki]	~	[bʊk]	"book-i" - "book"
	[wɔki]	~	[wɔk]	"rocky" - "rock"
Target [g]	[bægi]	~	[bæg]	"baggy" - "bag"
	[hʌgɪŋ]	~	[hʌg]	"hugging" - "hug"

Moreover, in intervocalic and final positions, a phonemic voice distinction was maintained among stops, as illustrated by the following minimal and near minimal pairs:

"soapy"	[do͜ʊpi]	-	[wo͜ʊbi]	"robe-i"
"cup"	[tʌp]	-	[dʌb]	"tub"
"because"	[ditɑ]	-	[didɚ]	"teacher"
"cut"	[dʌt]	-	[mʌd]	"mud"
"chicken"	[dɪkɨn]	-	[dɪgi]	"ziggy"
"back"	[bæk]	-	[bæg]	"bag"

In word-initial position, Becca also produced stops. However, voiced and voiceless stops were in free variation, as in the forms:

[p] ~ [b]	[pe͡ɪ]	~	[be͡ɪ]	"play"
	[pɪ̂ks]	~	[bɪ̂ks]	"fix"
	[pe͜ɪdʒi]	~	[be͜ɪːtʃ]	"page-i" - "page"
[t] ~ [d]	[tiː]	~	[di]	"see"
	[tɪd]	~	[dɪθ]	"kid(s)"
	[tʌmpɪn]	~	[dʌmpɪn]	"something"
[k] ~ [g]	[kɔf]	~	[gɔfɪŋ]	"cough" - "coughing"
	[ko͜ʊm]	~	[go͜ʊmi]	"comb" - "comb-i"

A voice contrast was evidently not maintained among word-initial stops. This was further illustrated by Becca's homophonous productions of potential minimal and near minimal pairs:

"pig"	[bɪg]	-	[bɪg]	"big"
"pie"	[ba͜ɪ]	-	[ba͜ɪt]	"bite"
"to"	[du]	-	[du]	"do"
"tear"	[dɪə]	-	[dɪə]	"deer"
"coat"	[go͜ʊt]	-	[go͜ʊt]	"goat"

These data indicate that Becca used voicing contrastively in the intervocalic and final positions of words; however, she did not use this contrast word-initially.

Thus, the results of the phonological analyses indicated that both Aaron and Becca exhibited a similar pattern of error with regard to their use of obstruent stops. Stops were produced in all word positions. However, a voice distinction was only maintained for post-vocalic stops; word-initially, voicing was not systematically contrastive for either child.

This particular property of the phonological systems of these two children is unusual in that their pattern of production is an apparent violation of a known substantive universal, the voice contrast hierarchy (Dinnsen & Eckman, 1975, 1978). This universal, which is based on phonological contrasts (and not strictly phonetic parameters), states that the presence of a voice contrast word-finally implies the contrast word-medially, and that in turn, implies the contrast word-initially. The universal hierarchy predicts that a language which maintains a voice contrast in any word position will also maintain the contrast in all implied positions, but not necessarily in implying positions. Initial position (least marked) is the most favored position for a voice contrast, and final position (most marked) is the least favored position for a voice contrast.

Typological evidence from primary (first acquired, natural) languages has indicated that there are languages with no voice contrast in any position, e.g., Korean.[6] There are languages with a voice contrast initially, but not intervocalically or finally, e.g., Corsican, Sardinian. There are languages with a voice contrast initially and intervocalically, but not finally, e.g., Polish, German. Of course, there are languages with a voice contrast in all positions, e.g., English. The universal excludes the possibility that a primary language will maintain a voice contrast in intervocalic and final positions without also maintaining this contrast word-initially. No known primary languages have shown patterns contrary to this observed relationship.

To date, there are no available data from normal language acquisition which bear upon this universal. While there have been numerous reports on the acquisition of voicing by children of different languages (Gilbert, 1977; Kewley-Port & Preston, 1974; Krause, 1982; Locke, 1983; Macken, 1980; Macken & Barton, 1977, 1980; Raphael, Dorman, & Geffner, 1980; Smith, 1978; Zlatin & Koenigsknecht, 1976), these investigations have only focused on the use of the voice contrast in one word position. Development of the voice contrast across word positions for individual children is an area of investigation which needs further attention.

Similarly, there are no available data from misarticulating children which bear upon this universal. The potential violation of this universal by the two children of this study serves as additional motivation for examining their disordered sound systems in greater detail. While phonological data indicate that these children did not produce a word-initial voice contrast, perhaps acoustic phonetic evidence would show that this contrast was being systematically marked, as has been observed in at least one stage of acquisition of the voice contrast by normally developing children (Barton & Macken, 1980; Macken & Barton, 1977, 1980). If an acoustic analysis revealed that a systematic distinction was being produced by the children (although not detected in the phonetic transcriptions), these data would be in agreement with the universal voice hierarchy. Given the apparent phonological similarity of the sound systems of these two children and given the apparent uniqueness of this error pattern, an acoustic phonetic study was designed to further investigate word-initial voicing.

Acoustic Phonetic Analyses

Data collection

A naming game was developed to elicit comparable utterances from each child for measurement purposes. Before the actual test session, each child participated individually in a pretraining session in order to instruct him or her in this game. The game required that the child embed the name of an object or picture in the carrier phrase, "Say _____ again." The game proceeded in a manner similar to "Simon Says," with pictures and objects presented in sequence and the child spontaneously producing the desired phrase. The child stayed "in" the game and earned points for saying the entire utterance with the embedded target word, e.g., "say sun again." The child was "out" of the game if only the target word was named, e.g., "sun," only part of the carrier phrase was produced, e.g., "sun again," or there was a pause between "say" and the target word, e.g., "say (pause) sun."

The actual test session proceeded in much the same manner as pretraining. Six minimal pairs were selected as test items for spontaneous production. Test items were common, picturable words familiar to the children. Two exemplars for each voiced and voiceless stop in each place of articulation were used: "pig" - "big," "peach" - "beach" (bilabial stops); "town" - "down," "tear" - "deer" (alveolar stops); "coat" - "goat," "curl" - "girl" (velar stops). These items were randomly presented to each child for production in the carrier phrase, "Say ____ again." Each test item was elicited 15 times, for a total sample size of 180 tokens per child. Samples were collected individually for each child and tape-recorded in a sound-insulated clinical treatment room over three consecutive days.

Data analysis

Wide-band spectrograms (300 Hz filter) with high frequency shaping were made on a Voice Identification Series 700 Sound Spectrograph. Measurements were made relative to the vowel in the carrier word "say" and the following stop and vowel in the test word. Measurements were made to the nearest 5 milliseconds (msec) for two different timing intervals: stop closure duration and voice onset time (VOT). Stop closure duration was defined as the interval from the offset of periodic vertical striations in the first and second formants of the vowel in the word "say" to the sudden spiked vertical increase in amplitude, indicating a stop release burst. VOT was defined as the interval from the stop release burst (as above) to the onset of periodic vertical striations in the following first formant. These particular parameters were selected for measurement since they have been cited as two cues to the voice distinction in word-initial stops (Delattre, Liberman, & Cooper, 1955; Flege & Port, 1980; Lisker & Abramson, 1964, 1967; Malecot, 1968; Stathopoulos & Weismer, 1983; Zlatin, 1974). In English, closure duration may not be a primary cue to the voiced-voiceless contrast in prestressed stops (for review of relevant literature, see Flege & Brown, 1982). However, closure duration is a cue to voicing in prevocalic stops in other languages (e.g., Arabic). Since children may produce phonetic distinctions which are unlike those used by adults, it is plausible that closure duration functioned as a cue to the voice contrast for the two children of this study.

For each child, several measurements had to be discarded from the original data set. The measurements were discarded because of poor recordings including a child's production being spoken too softly or with extraneous noises, such as hand clapping or table tapping; or because the closure duration exceeded 250 msec, indicative of a pause between the carrier word, "say" and the following test word. In any of these cases, measurement of the noted parameters would prove unreliable (Maxwell, 1981a, 1981b; Maxwell & Weismer, 1982). This resulted in approximately 16% of the 180 tokens for Aaron being discarded; 152 tokens were subjected to analysis. For Becca, approximately 37% of the tokens had to be discarded; 114 tokens were suitable for analysis.

Analyses of variance were calculated for closure duration and VOT for each child. A 2 (voice) x 3 (place of articulation) unbalanced factorial design was used. The criterion for significance was set at $p < .025$ for each comparison.

Intrajudge reliability

An estimate of measurement reliability was calculated for 10% of each child's tokens used in the acoustic analysis. Spectrograms were remeasured by the first author approximately nine months after the original measurements were obtained. Intrajudge reliability was determined by calculating a mean difference score (msec) between the initial measurements and the remeasurements (cf. Charles-Luce, 1985) for both closure duration and VOT. For Aaron, the mean difference scores were +/- 5.67 msec for closure duration and +/- .67 msec for VOT. For Becca, mean difference scores were +/- 6.67 msec for closure duration and +/- 4.58 msec for VOT.

Results

Aaron's acoustic analysis. Mean values for closure duration and VOT for voiced and voiceless stops in each place of articulation are presented in Figures 1 and 2; means and standard deviations are reported in Table 1.

Insert Figures 1 and 2 about here

Insert Table 1 about here

For closure duration, the results of the analysis of variance indicate no significant main effect for voice [F(1,146)=1.34] or interaction between voice and place of articulation [F(2,146)=1.00]. There was, however, a significant main effect for place of articulation [F(2,146)=16.65, p<.025]. Closure duration values for place of articulation generally followed the sequence noted for adults (Flege & Port, 1980; Stathopoulos & Weismer, 1983), with bilabials being of longer duration than alveolars or velars.

The results of the analysis of variance for VOT also indicated no significant main effect for voice [F(1,146)=2.05] or interaction between voice and place [F(2,146)=2.67]. There was a significant main effect for place of articulation [F(2,146)=13.57, p<.025]. Again, place trends approximated those reported for adults (Klatt, 1975; Lisker & Abramson, 1964, 1967; Port & Rotunno, 1979); i.e., velars have longer VOT values than alveolars which, in turn, have longer VOT values than bilabials.

The results of the acoustic analysis for Aaron indicated neither closure duration nor VOT were used to mark a voice distinction among word-initial stops. It may have been the case, however, that this child used another parameter, such as amplitude of the burst or fundamental frequency of the following vowel, to achieve the voice contrast. Methodologically, of course, it would have been impossible to rule out all potentially relevant parameters. Thus, these findings, while specific to the acoustic parameters measured, do support the phonological description indicating the absence of a word-initial voice contrast in stops.

Becca's acoustic analysis. Mean values for closure duration and VOT for voiced and voiceless stops in each place of articulation are displayed in Figures 3 and 4; means and standard deviations are also reported in Table 2.

Insert Figures 3 and 4 about here

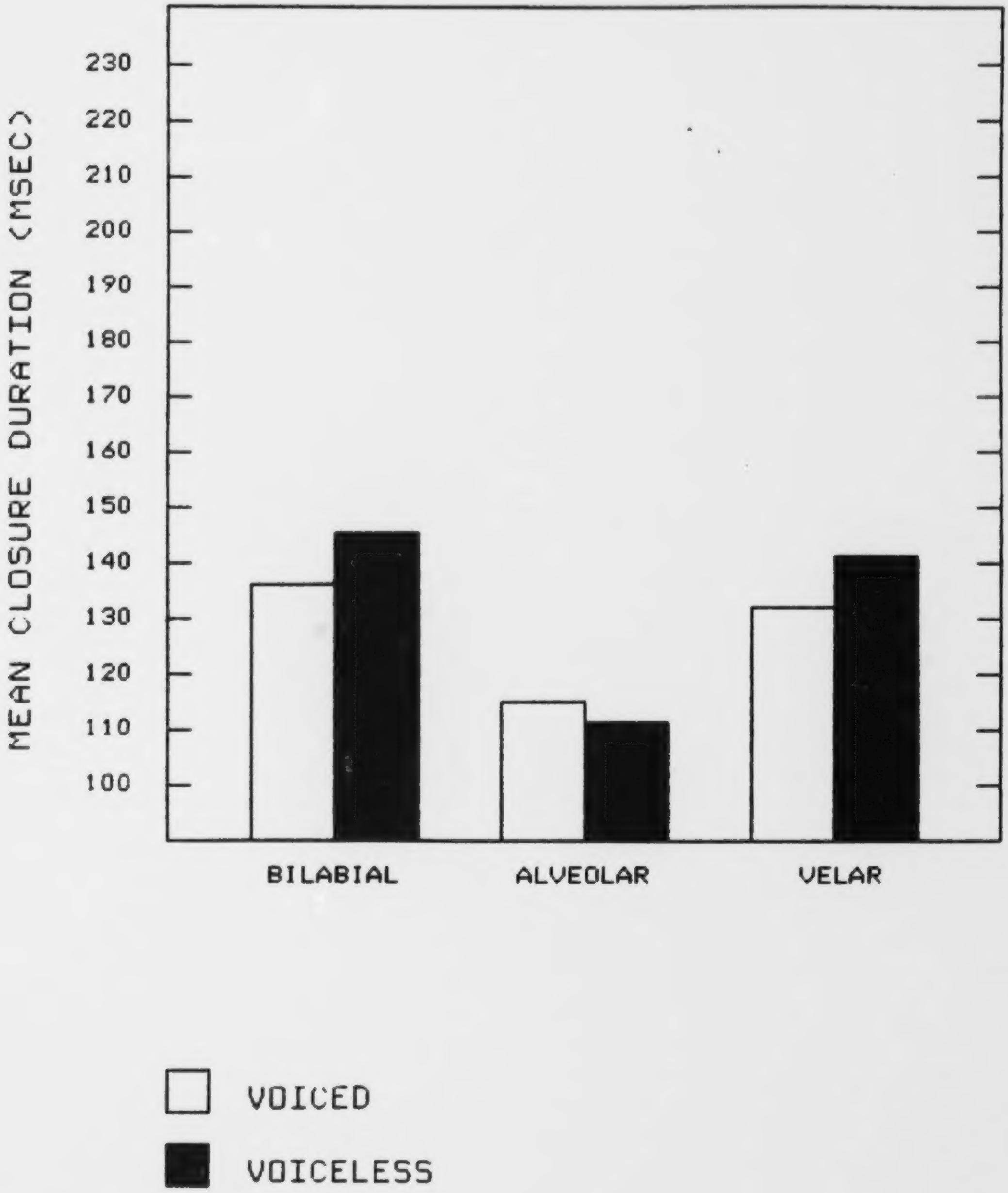


Figure 1. Mean closure duration values (msec) for voiced and voiceless stops in each place of articulation for Aaron.

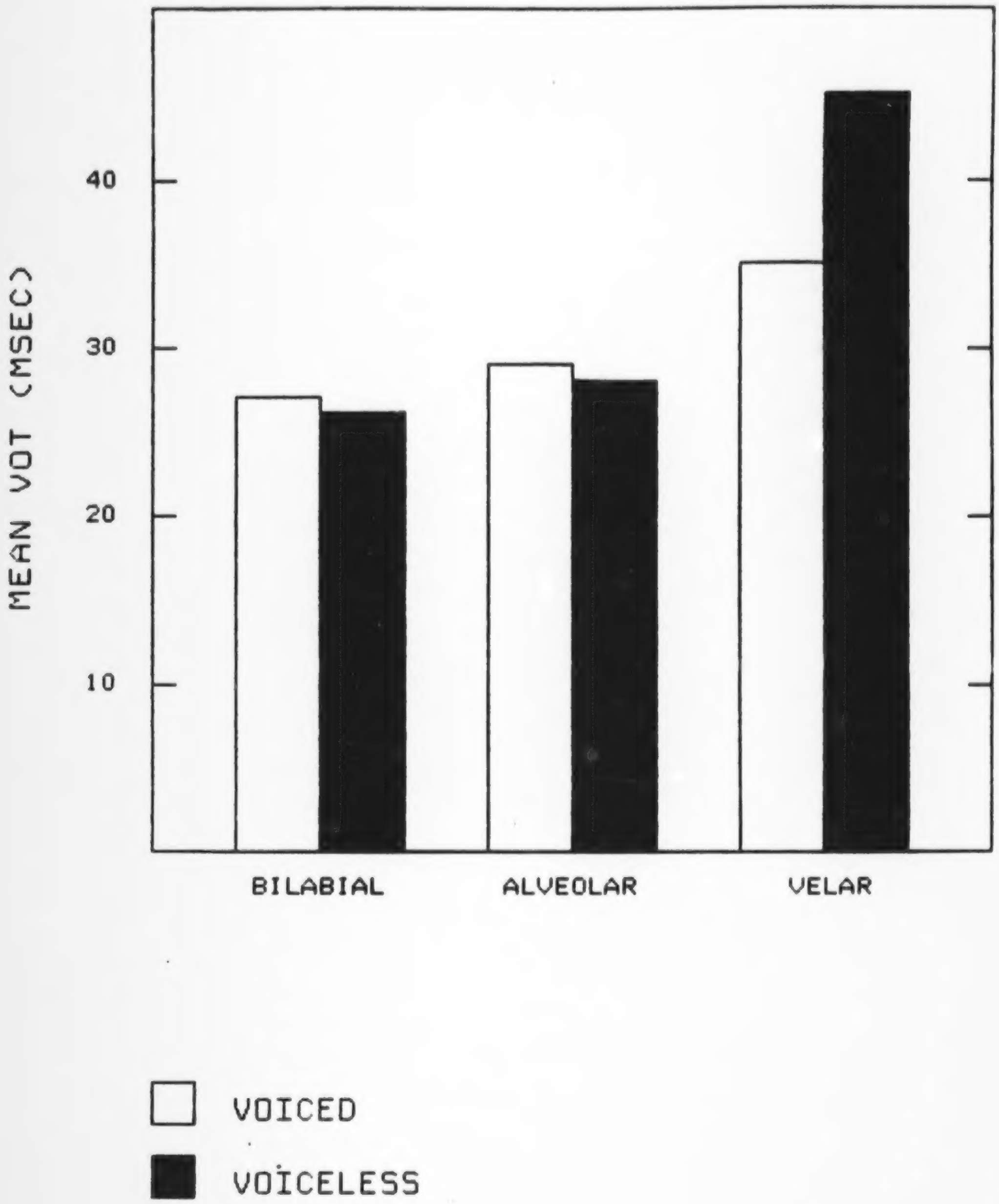


Figure 2. Mean VOT values (msec) for voiced and voiceless stops in each place of articulation for Aaron.

Table 1. Means and standard deviations for closure duration and VOT in msec for each place of articulation for Aaron.

Place of Articulation		Closure Duration			VOT		
		n	$\bar{x}$	σ	n	$\bar{x}$	σ
Bilabials	+voice	22	135.68	20.08	22	26.59	10.28
	-voice	29	144.66	31.05	29	25.69	9.33
Alveolars	+voice	25	114.80	15.84	25	28.60	15.91
	-voice	29	111.38	25.14	29	28.45	14.09
Velars	+voice	25	131.60	24.53	25	34.60	11.98
	-voice	22	140.68	33.64	22	45.45	20.52

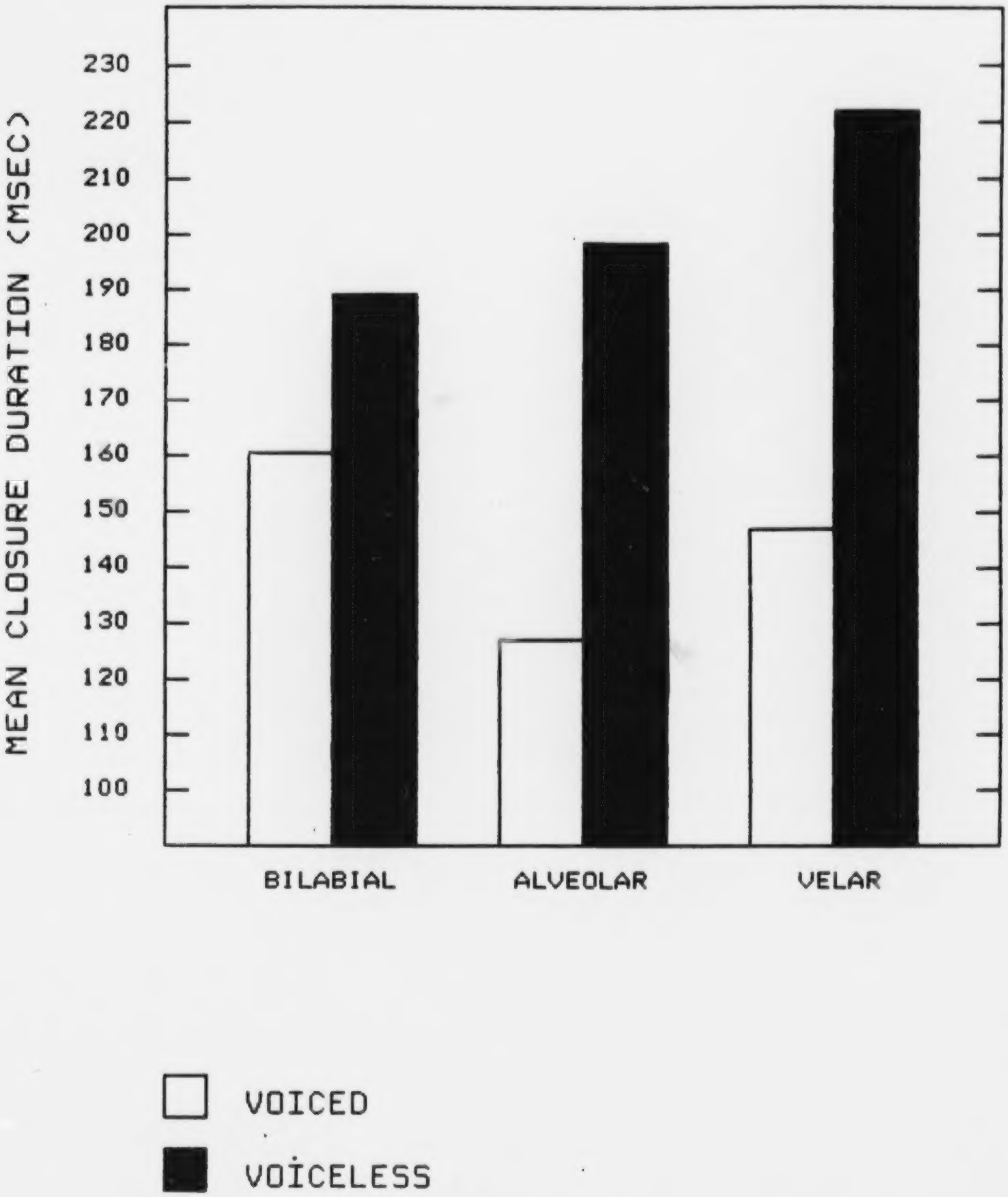


Figure 3. Mean closure duration values (msec) for voiced and voiceless stops in each place of articulation for Becca.

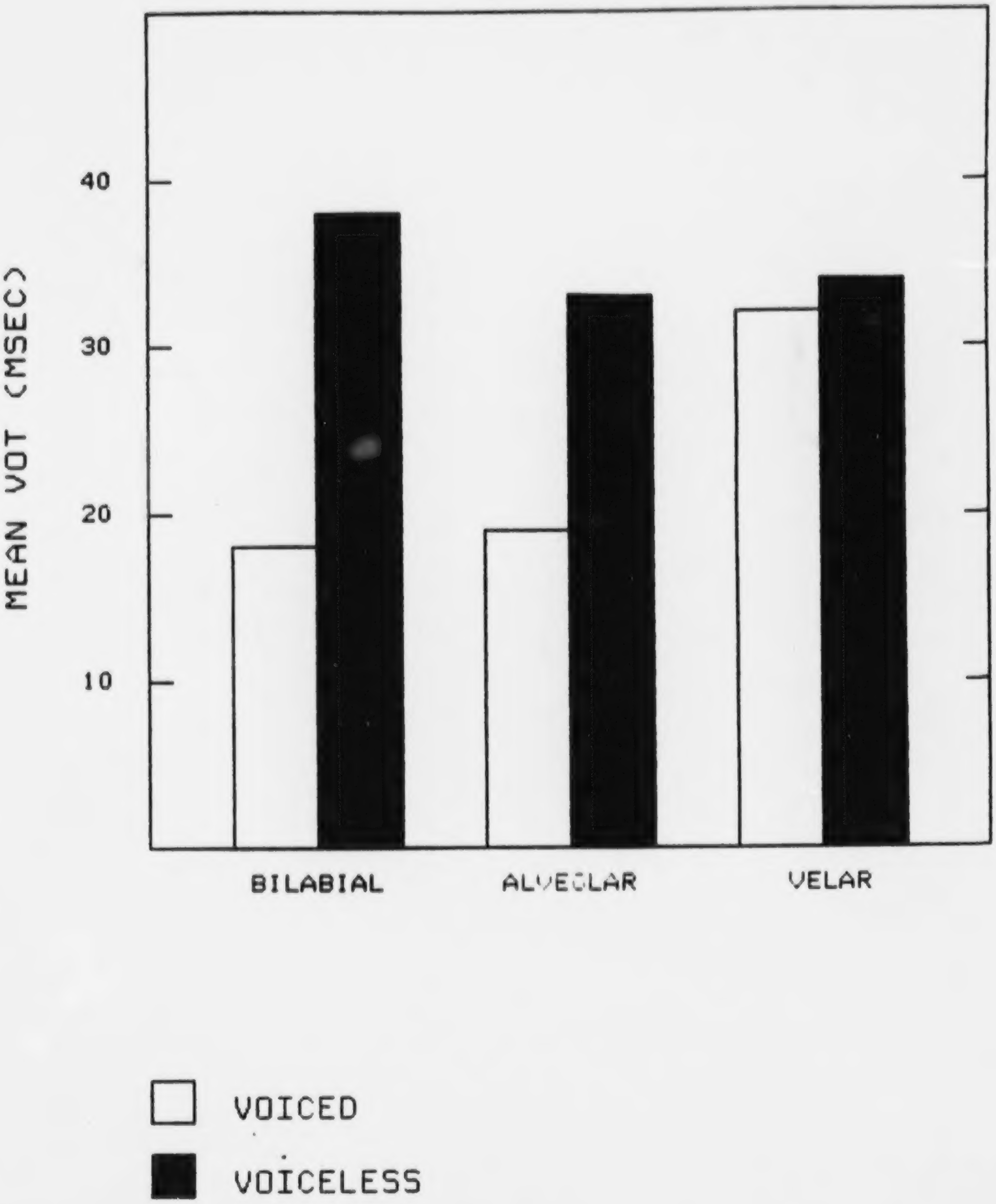


Figure 4. Mean VOT values (msec) for voiced and voiceless stops in each place of articulation for Becca.

Insert Table 2 about here

For closure duration, results of the analysis of variance indicated that there was a significant voice distinction among stops [F(1,108)=58.39, p<.025], with voiceless stops of greater duration than voiced stops. This finding is generally consistent with closure duration data reported for adults (Flege & Port, 1980; Malecot, 1968; Stathopoulos & Weismer, 1983). There was no significant main effect for place of articulation [F(2,108)=2.58]; the relative place sequence (bilabial > alveolar or velar) was maintained, but only for voiced stops. A significant voice by place interaction [F(2,108)=4.23, p<.025] was noted; that is, mean voiced-voiceless distinctions were greater for alveolar and velar stops than bilabial stops.

For VOT, the results of the analysis of variance indicated a significant main effect for voice [F(1,108)=13.65, p<.025]; VOT values for voiceless stops were of greater duration than for voiced stops. There was no significant main effect for place of articulation [F(2,108)=1.55]; however, relative place trends (velar > alveolar > bilabial) were maintained, but only for voiced stops. There was no significant interaction between voice and place [F(2,108)=2.31].

The results of the acoustic analysis for Becca indicated that both timing intervals, closure duration and VOT, were used to affect a voice contrast among word-initial stops. These results, while specific to the test items and acoustic parameters measured, were not consistent with those of the phonological analysis which failed to identify a distinction in voice for word-initial stops. Becca demonstrated more productive knowledge of the voice contrast in word-initial stops than was originally determined by the phonological analysis.

Discussion

These two children displayed superficially similar patterns of error, but their productive knowledge of the voice contrast in stops was different. For Aaron, phonological and acoustic phonetic data were converging. Given all the available data, this child did not produce or acoustically mark the voice distinction word-initially. For Becca, on the other hand, phonological and acoustic data were non-converging. Taken together, related sources of data indicated that Becca was, in fact, producing a voice contrast for stops in all positions; word-initially, however, this contrast was not perceptible to adult listeners. Moreover, acoustic phonetic data were necessary to accurately characterize Becca's productive knowledge of the voice contrast.

Differences in productive knowledge of the word-initial voice contrast may also have implications for treatment goals. For Becca, the voice contrast may not need to be taught since she was already consistently producing the relevant distinction. Her problem would appear to be a matter of phonetic implementation, and her error would be viewed as phonetic in nature. For Aaron, it is clear that his error is phonological in nature and requires learning the voice contrast in initial position. It is less clear, however, what the goals of treatment should be in this case. The most obvious recommendation is that the voice contrast should be directly taught since it

Table 2. Means and standard deviations for closure duration and VOT in msec for each place of articulation for Becca.

Place of Articulation		Closure Duration			VOT		
		n	$\bar{x}$	σ	n	$\bar{x}$	σ
Bilabials	+voice	25	160.00	44.21	25	18.40	10.77
	-voice	18	188.61	37.57	18	37.78	24.69
Alveolars	+voice	23	126.74	44.05	23	19.13	9.73
	-voice	15	198.00	35.24	15	33.33	18.19
Velars	+voice	24	146.67	31.23	24	32.29	17.38
	-voice	9	222.22	25.99	9	34.44	13.10

was not systematically produced in word-initial position. Here, treatment might take the form of minimal pair contrast training (cf. Ferrier & Davis, 1973) among word-initial voiced and voiceless stops. In light of some recent research findings (Dinnsen & Elbert, 1984), however, it may not be necessary to teach this contrast. Specifically, Dinnsen and Elbert demonstrated that if a child is taught to produce more marked aspects of phonology, the acquisition of unmarked aspects of phonology will occur without direct treatment. For Aaron, there is some likelihood that voicing in the unmarked word-initial position would be spontaneously acquired, since he already produced this contrast in more marked post-vocalic positions.[7] In this case, production of the word-initial voice contrast might only need to be monitored or minimally treated in the course of clinical intervention. Thus far, there are no available data to suggest which of these two treatment goals is to be preferred, but these considerations are suggestive of future research.

The results of the acoustic and phonological analyses also bear upon the accuracy of the voice contrast hierarchy. By examining related sources of data, it is possible to establish whether the sound systems of these children violate the voice contrast hierarchy. In the case of Becca, even though the phonological analysis did not converge with the acoustic evidence, the acoustic evidence did indicate that the voice contrast hierarchy was maintained. This child produced a voice distinction in both marked and unmarked word positions; however, the strength of the contrast varied by position. That is, in post-vocalic positions, the contrast was perceptible to listeners; whereas, word-initially, it was not.

In the case of Aaron, the voice contrast hierarchy was not obviously maintained. This child's sound system violated the linguistic universal and, therefore, was not like any other known phonological system. Evidence of this type can be brought to bear on the formulation of the voice contrast hierarchy. It may be necessary to revise this universal to accommodate this child's unusual phonological system. While it is not uncommon to find well-defined exceptions to language universals (cf. Gamkrelidze, 1975), qualifications of this particular universal are not readily apparent. More likely, this type of evidence also may bear on the characterization of Aaron's sound system. It has been suggested (Connell, 1982) that violations of language universals may serve as a means of systematically determining the severity of a child's speech sound disorder. Aaron's violation of the voice contrast hierarchy could be taken as evidence that his sound system (or at least part of the system) was structurally unlike that of other languages, indicative of a "severe" phonological disorder.[8]

In conclusion, the results of this study demonstrated that: (1) phonological and acoustic phonetic sources of data were necessary to accurately describe the errored sound systems of two children; (2) despite superficially similar phonological patterns, the children had very different productive knowledge of the relevant voice contrast; and (3) differences in productive knowledge may be reflected in subsequent treatment goals. These findings were consistent with those of previous research (Maxwell, 1981a, 1981b; Maxwell & Weismer, 1982; Weismer et al., 1981), which underscored the importance of bringing phonological and acoustic phonetic data to bear upon the clinical assessment and treatment of children with phonological disorders. What was unique about this study, however, was that related sources of evidence were also used to empirically validate a substantive universal, the voice contrast hierarchy. It was only possible to establish "true" violations of the voice contrast hierarchy when converging sources of data, phonological and acoustic phonetic, were examined. Counterexamples to linguistic universals, when based on both phonological and acoustic phonetic evidence, may aid in defining the full range and nature of language types, in both normal and phonologically disordered speakers.

Endnotes

1 Another possible interpretation of these results, that was not considered by Weismer et al. (1981), is that the children demonstrated knowledge of a phonemic vowel length distinction rather than knowledge of final obstruents. This interpretation, however, would have other theoretical consequences that are problematic, i.e., phonemes with defective distributions.

2 The claim that a phonological rule deleted final obstruents was further motivated by morphophonemic alternations for these two children but not for the third child.

3 The underlying representation of the morpheme "rub" is /wæb/; "cob" is represented as /kɔb/. An optional rule of word-final devoicing applied in this child's sound system; thus, the phonetic forms of these morphemes do not have final voiced segments, resulting in the productions, [wæp] and [kɔp].

4 The "voice contrast" or "voicing," as used throughout this paper, refers to a phonological distinction among voiced and voiceless obstruents. This phonological voice distinction is, of course, implemented phonetically in a variety of ways in various contexts in different languages.

5 The alternation between [ɾ] and [t] is consistent with adult productions. This child, like adult speakers of English, used a rule of intervocalic flapping.

6 Superficially, it may appear that Korean is not an appropriate example of this point since both voiced and voiceless obstruents occur between vowels. The voiced obstruents, however, correspond to and alternate with lax voiceless unaspirated stops in other positions and are thus entirely predictable by an allophonic rule. Phonologically all stops in Korean are described as voiceless, i.e., tense and lax unaspirated stops and aspirated stops (Martin, 1951; Moon, 1974).

7 Of course, this does not explain why the child does not already produce the voice contrast in initial position given that he produces it in more marked contexts. Claims about markedness do not in all cases correspond with claims about order of acquisition (Locke, 1983).

8 Aaron was enrolled in a clinical research program subsequent to his participation in this study. In this intervention program, other errored aspects of his sound system were targeted for treatment. After approximately nine months of intervention, Aaron produced targeted sounds with only 50% accuracy. Both the level of performance following treatment and the length of time enrolled in treatment suggest that Aaron's sound system may, in fact, have been "severely" disordered.

References

Barton, D., & Macken, M. (1980). An instrumental analysis of the English voicing contrast in four-year-olds. Language and Speech, 23, 159-169.

Charles-Luce, J. (1985). Word-final devoicing in German: Effects of phonetic and sentential contexts. Journal of Phonetics, 13, 309-324.

Connell, P.J. (1982). Markedness differences in the substitutions of normal and misarticulating children. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Toronto, Canada.

Delattre, P.C., Liberman, A.M., & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 27, 769-773.

Dinnsen, D.A. (1984). Methods and empirical issues in analyzing functional misarticulation. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), Phonological theory and the misarticulating child (ASHA Monographs No. 22, pp. 5-17). Rockville, MD: ASHA.

Dinnsen, D.A. (1985). A re-examination of phonological neutralization. Journal of Linguistics, 21, 265-279.

Dinnsen, D.A., & Eckman, F. (1975). A functional explanation of some phonological typologies. In R. Grossman, J. San, & T. Vance (Eds.), Functionalism (pp. 126-134). Chicago: Chicago Linguistic Society.

Dinnsen, D.A., & Eckman, F. (1978). Some substantive universal in atomic phonology. Lingua, 45, 1-14.

Dinnsen, D.A., & Elbert, M. (1984). On the relationship between phonology and learning. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), Phonological theory and the misarticulating child (ASHA Monographs No. 22, pp. 59-68). Rockville, MD: ASHA.

Ferrier, E., & Davis, M. (1973). A lexical approach to the remediation of sound omissions. Journal of Speech and Hearing Disorders, 38, 126-130.

Flege, J.E., & Brown, W.S. (1982). Effects of utterance position on English speech timing. Phonetica, 39, 337-357.

Flege, J.E., & Port, R.F. (1980). Cross-language phonetic interference from Arabic to English. Research in phonetics (Report No. 1, pp. 99-136). Bloomington, IN: Department of Linguistics, Indiana University.

Gamkrelidze, T.V. (1975). On the correlation of stops and fricatives in a phonological system. Lingua, 35, 231-261.

Gilbert, J.H.V. (1977). A voice onset time analysis of apical stop production in 3-year olds. Journal of Child Language, 4, 103-110.

Goldman, R., & Fristoe, M. (1969). Goldman-Fristoe test of articulation. Circle Pines, MN: American Guidance Service.

Hoffman, P.R., Stager, S., & Daniloff, R.G. (1983). Perception and production of misarticulated /r/. *Journal of Speech and Hearing Disorders*, *48*, 210-215.

Kenstowicz, M., & Kisseberth, C. (1979). *Generative phonology.* New York: Academic Press.

Kewley-Port, D., & Preston, M.S. (1974). Early apical stop production: A voice-onset time analysis. *Journal of Phonetics*, *2*, 195-210.

Klatt, D.H. (1975). Voice-onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*, 687-703.

Kornfeld, J.R., & Goehl, H. (1974). A new twist to an old observation: Kids know more than they say. *Papers from the Parasession on Natural Phonology* (Chicago Linguistic Society), *10*, 210-219.

Krause, S.E. (1982). Vowel duration as a cue to postvocalic consonant voicing in young children and adults. *Journal of the Acoustical Society of America*, *71*, 990-995.

Lisker, L., & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384-422.

Lisker, L., & Abramson, A.S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*, 1-28.

Locke, J.L. (1983). *Phonological acquisition and change*. New York: Academic Press.

Macken, M. (1980). Aspects of the acquisition of stop systems: A cross-linguistic perspective. In G.H. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child phonology: Production* (Vol. 1, pp. 143-168). New York: Academic Press.

Macken, M., & Barton, D. (1977). A longitudinal study of the acquisition of the voicing contrast in American-English word-initial stops, as measured by voice-onset time. *Papers and Reports in Child Language Development*, *14*, 74-120.

Macken, M., & Barton, D. (1980). A longitudinal study of the voicing contrast in American English word-initial stops, as measured by voice onset time. *Journal of Child Language*, *7*, 41-74.

Malecot, A. (1968). The force of articulation of American stops and fricatives as a function of position. *Phonetica*, *18*, 95-102.

Martin, S. E. (1951). Korean phonemics. *Language*, *27*, 519-533.

Maxwell, E.M. (1981a). The use of acoustic phonetics in phonological analysis. *Journal of the National Student Speech, Language, and Hearing Association*, *9*, 20-37.

Maxwell, E.M. (1981b). *A study of misarticulation from a linguistic perspective*. Doctoral dissertation, Indiana University, Bloomington, IN. (Also distributed by the Indiana University Linguistics Club, Bloomington, IN)

Maxwell, E.M., & Rockman, B.K. (1984). Procedures for linguistic analysis of misarticulated speech. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monographs No. 22, pp. 69-84). Rockville, MD: ASHA.

Maxwell, E.M., & Weismer, G. (1982). The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics*, *3*, 29-43.

Menn, L. (1983). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.), *Language production* (Vol. 2, pp. 3-50). New York:Academic Press.

Menyuk, P. (1972). Clusters as single underlying consonants: Evidence from children's production. In A. Rigault, & R. Charbonneau (Eds.), *Proceedings of the seventh international congress of phonetic sciences* (pp. 1161-1165). The Hague: Mouton.

Moon, Y. S. (1974). *A phonological history of Korean*. Unpublished doctoral dissertation, University of Texas, Austin.

Port, R.F., & Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America*, *66*, 654-662.

Raphael, L.J., Dorman, M.F., & Geffner, D. (1980). Voicing-conditioned durational differences in vowels and consonants in the speech of three- and four-year-old children. *Journal of Phonetics*, *8*, 335-342.

Smith, B.L. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics*, *6*, 37-68.

Stathopoulos, E.T., & Weismer, G. (1983). Closure duration of stop consonants. *Journal of Phonetics*, *11*, 395-400.

Straight, H.S. (1980). Auditory versus articulatory phonological processes and their development in children. In G.H. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child phonology: Perception* (Vol. 2, pp. 43-71). New York: Academic Press.

Weismer, G. (1984). Acoustic analysis strategies for the refinement of phonological analysis. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monographs No. 22, pp. 30-52). Rockville, MD: ASHA.

Weismer, G., Dinnsen, D.A., & Elbert, M. (1981). A study of the voicing distinction associated with omitted word-final stops. *Journal of Speech and Hearing Disorders*, *46*, 320-327.

Zlatin, M. (1974). Voicing contrast: Perceptual and productive voice onset time characteristics of adults. Journal of the Acoustical Society of America, 56, 981-994.

Zlatin, M.A., & Koenigsknecht, R.A. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. Journal of Speech and Hearing Research, 19, 93-111.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

On the Assessment of Productive Phonological Knowledge*

Judith A. Gierut

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*This research was supported, in part, by NIH Training Grant NS-07134 to Indiana University in Bloomington. I would like to thank Daniel Dinnsen, Ronald Gillam, Kathy Hoyt, and Michael Smith for their comments and suggestions on an earlier version of this manuscript.

Abstract

This paper describes a new conceptual framework that has been introduced in the literature (cf. Elbert, Dinnsen, & Weismer, 1984; Gierut, 1985) for the assessment of functional speech sound disorders in children. Specifically, the concept of productive phonological knowledge is defined and described relative to disordered sound systems. It is further illustrated how productive phonological knowledge may be used in the clinical assessment and treatment of speech sound errors.

On the Assessment of Productive Phonological Knowledge

One of the keys to a successful and effective clinical treatment program is an accurate assessment and diagnosis of the speech disorder. In the assessment process, we establish what the speaker already knows about the language and also, what he or she has yet to learn in order to become a proficient and intelligible user of the language. The subsequent diagnosis of the disorder leads to the identification and selection of appropriate treatment goals. Our treatment programs follow directly from the assessment and diagnosis of the speech disorder and, therefore, can only be as effective as the assessment is thorough and the diagnosis is accurate. This paper examines the assessment process with regard to phonological disorders and presents a new conceptual framework for describing and evaluating speech sound problems in children.

Assessing phonological disorders in children usually involves identifying the *pattern* underlying a child's error productions (cf. Bernthal & Bankson, 1981). There are several different frameworks, e.g., place-voice-manner analysis (Turton, 1973), distinctive feature analysis (McReynolds & Engmann, 1975), phonological process analysis (Ingram, 1981; Shriberg & Kwiatkowski, 1980), that may be used to determine a child's pattern of production. Each framework provides a characterization of the systematicity and consistency of a child's error productions; the goals of treatment focus on interrupting the observed pattern of error.

Recently, a new framework for assessing speech sound disorders has been introduced (cf. Elbert, Dinnsen, & Weismer, 1984; Gierut, 1985). This framework lies within theoretical linguistics and is known as standard generative phonology (Chomsky & Halle, 1969; Kenstowicz & Kisseberth, 1979). Standard generative phonology is a *qualitative*, rather than quantitative, description of the structure and function of sounds in a language. That is, generative phonology identifies *which* sounds are present in a speaker's phonological system, but also indicates *how* these sounds are used. Generative phonology was initially developed as a tool for evaluating and describing what adult speakers of primary languages, such as French or German, know about the sound system of their language. The procedures of generative phonology are also applicable to applied disciplines as speech-language pathology and second language instruction to evaluate what a speaker knows about the sound system of the language being learned.

Generative phonology, like other frameworks for assessing disordered sound systems, identifies the pattern governing a child's error productions. Generative phonology characterizes the error pattern by assessing both a child's *performance* and *competence* (cf. Foss & Hakes, 1978).[1] The assessment of performance provides a surface level evaluation of the sounds that are produced in error; the assessment of competence provides an evaluation of the underlying use and function of sounds. Together, these features characterize a child's *productive phonological knowledge* of the target sound system. It is the evaluation of a child's productive phonological knowledge that makes this framework unique and different from other approaches to assessing speech sound disorders.

Also, like other frameworks, generative phonology and the assessment of productive phonological knowledge lead to the selection of appropriate treatment goals. That is, a child's productive phonological knowledge can be used in a structured way to plan and implement treatment programs (Gierut, 1985; Gierut, Dinnsen, & Elbert, 1984; Gierut & Elbert, 1983, 1985; Gierut, Elbert, & Dinnsen, in review). Specifically, a child's productive phonological knowledge can be ranked on a continuum ranging from "most" to "least" knowledge relative to the (adult) target. Sounds for treatment can be selected directly from the continuum of knowledge. Sounds of which a child has "most" knowledge will be easier to learn than sounds of which a child has "least" knowledge (Dinnsen & Elbert, 1984; Elbert, Dinnsen, & Powell, 1984; Gierut, 1985; Rockman, 1983). Moreover, when productive phonological knowledge is assessed prior to treatment, predictions can be made about a child's learning during treatment; the relative success of a treatment program can be estimated a priori. A child's productive phonological knowledge, then, is closely linked to learning during treatment. This relationship between productive phonological knowledge and learning is unique to this assessment framework; the knowledge-learning relationship has not been observed when other approaches to evaluating speech sound disorders have been used.

Thus, the framework of generative phonology bears similarity to other frameworks used in assessing speech sound disorders. Generative phonology identifies the pattern underlying error productions and aids in the selection of suitable treatment goals. Generative phonology, however, is unique in that it provides an assessment of a child's productive phonological knowledge, including both performance and competence. It also allows for predictions to be made about a child's learning during treatment based upon productive phonological knowledge.

The purpose of this paper is to introduce certain aspects of this new conceptual framework for assessment by describing productive phonological knowledge and by demonstrating how productive phonological knowledge can be used in the clinic. The paper is organized with a definition of phonological knowledge presented first, followed by a description of the six types of phonological knowledge that speech disordered children display. The sound systems of two phonologically disordered children are then examined to illustrate how productive phonological knowledge can be applied in clinical assessment and treatment.

Productive Phonological Knowledge

What is phonological knowledge?

Productive phonological knowledge has been defined (Dinnsen, 1984; Dinnsen & Elbert, 1984; Elbert & Gierut, 1986) as the idiosyncratic, unpredictable properties of productive language that are learned and stored in a speaker's lexicon. Phonological knowledge refers to those aspects of production and properties of the sound system that are specific to a particular language. These properties must be learned and cannot be generated by rule. Consider the following example. In English, the sounds [d], [ɔ], [g] have no inherent meaning in and of themselves. When these sounds combine as [dɔg], however, they are associated with the morpheme, "dog," meaning "canine." This combination of sounds meaning "canine" is idiosyncratic and specific to English; other languages use other sound combinations to signal the same meaning (e.g., French [ʃiɛn], Spanish [perro]). This sequence of sounds meaning "canine" is also unpredictable. That is, there is no a priori or independent reason for the morpheme "dog" to be composed of three segments, for the segments to combine in a consonant-vowel-consonant sequence, for both consonants to be voiced stops, or for the first segment to be [d] versus any other sound. In other words, any other combination of sounds would be just as likely to signal the meaning "canine" in English as [dɔg]. This information about the morpheme "dog" is, therefore, idiosyncratic and unpredictable and must be learned by every speaker of English. These learned aspects of productive language are then stored, or represented, in a speaker's lexicon. This information constitutes a speaker's lexical or underlying representation of morphemes.

In addition to the idiosyncratic and unpredictable properties of productive language, phonological knowledge also refers to the rules associating sound to meaning; this aspect of phonological knowledge is predictable. Consider the case of plurals. There are three possible ways the plural morpheme may be pronounced in English, [s], [əz], or [z]. The plural is pronounced as [s] when preceded by a voiceless segment, as in "books" or "bats." It is realized as [əz] when preceded by [s,z,ʃ,ʒ,tʃ,ʤ], as in "buses" or "bushes." The plural is also pronounced as [z] when preceded by a voiced segment, as in "bees" or "bags." Note that each way of signalling the plural morpheme depends upon phonetic context. The different pronunciations of the plural morpheme are predictable from the phonetic context and, therefore, can be generated by rule. There is still, however, only one meaning associated with the morpheme, but there are three ways of signalling that meaning. A speaker learns the meaning of the morpheme and then generates the alternate pronunciations in the various contexts by a rule. This information constitutes the speaker's use of phonological rules. Together, the lexical representation of morphemes and the use of phonological rules describe a speaker's competence (or tacit knowledge) of the target sound system.

As mentioned, productive phonological knowledge also includes a characterization of a child's performance. This aspect of phonological knowledge specifies the sounds that are used by a child whether these are used correctly or not; i.e., the phonetic inventory. Those sounds that a child uses to contrast meaning are also identified; i.e., the phonemic inventory. Finally, the distribution of sounds in the child's sound system is noted. The distribution of sounds describes where sounds are used; i.e., whether a particular sound is used in all word positions and whether a sound is used for all target morphemes. The phonetic and phonemic inventories and the distribution of sounds describe a speaker's performance (or explicit knowledge) of the target sound system.

To summarize, productive phonological knowledge refers to a speaker's competence about the target sound system, including the unpredictable (i.e., lexical representation of morphemes) as well as the predictable (i.e., phonological rules) aspects of productive language. Productive phonological knowledge also refers to a speaker's performance of the target sound system, including the inventory and distribution of sounds.

Types of phonological knowledge

There are six different types of productive phonological knowlege that have been identified in phonologically disordered children thus far (Gierut, 1985). The six knowledge types have been observed both within and across children. The different types of knowledge emerge when the structure and function of sounds in a child's sound system are examined. In particular, three factors are taken into account: (1) the nature of a child's lexical representation of morphemes, either adult-like or nonadult-like, (2) the breadth of the distribution of sounds, extending either to some or all word positions or to some or all target morphemes, and (3) the use of phonological rules. Not coincidentally, these three factors are precisely those components which constitute a child's productive phonological knowledge. The six types of productive phonological knowledge are displayed in Table 1; a description and examples of each type of knowledge are presented in Table 2 (adapted from

Gierut et al., in review).

Insert Tables 1 and 2 about here

Notice, with reference to these tables, that sounds which are always produced correctly relative to the (adult) target are characterized as Type 1 knowledge. At the opposite extreme, sounds which are always produced incorrectly relative to the target are represented as Type 6 knowledge. Knowledge types 2 through 5 describe variations in sound production ranging between completely correct productions and completely incorrect productions relative to the adult. The six knowledge types, therefore, describe all logical combinations of the three factors: lexical representation, breadth of distribution, and use of rules. The different knowledge types capture not only the consistency, but also the full range of variability and inconsistency, that may be observed in a child's errored productions.

Assessing Productive Phonological Knowledge

There are four steps involved in the assessment of productive phonological knowledge: (1) obtaining a representative sample of speech, (2) describing a child's productive phonological knowledge of target sounds, (3) ranking productive phonological knowledge on a continuum, and (4) selecting treatment goals and sounds directly from the knowledge continuum. A brief description of each step follows; for a more complete discussion, the reader is referred to Elbert and Gierut (1986).

Obtaining a representative sample of speech

To determine a child's productive phonological knowledge, a representative speech sample must be obtained. Ideally, the sample should include both connected speech and spontaneous single word utterances. The sample should also meet the following criteria: (1) sample all target English sounds, (2) sample each sound in at least three word positions (initial, medial, and final), (3) sample each sound in each position in more than one word, and (4) sample each word more than one time. In addition, the sample should provide an opportunity for a child to produce potential minimal pairs and morphophonemic alternations. Minimal pairs (e.g., "pat"-"bat" or "cap"-"cab") provide information about those sounds a child is using contrastively as phonemes. Morphophonemic alternations are pairs such as "pig"-"piggie" or "miss"-"missing." Morphophonemic alternations sample a sound (e.g., [g], [s]) in a single morpheme (e.g., "pig", "miss") placed in different phonetic environments (e.g., intervocalic versus final position). Morphophonemic alternations provide information about the lexical representation of morphemes and the application of phonological rules. To date, there are two available protocols which meet these criteria (Gierut, 1985; Maxwell & Rockman, 1984).

Describing productive phonological knowledge

Having obtained a representative sample of speech, the second step in the assessment of productive phonological knowledge involves determining the type of knowledge that a child displays for each target sound. That is, each

KNOWLEDGE TYPES	LEXICAL REPRESENTATION	BREADTH OF DISTRIBUTION		PHONOLOGICAL RULE ACCOUNT
		Positions	Morphemes	
1	Adult-like	All	All	None
2	Adult-like	All	All	Optional or obligatory rules
3	Adult-like	All	Some	Fossilized forms
4	Adult-like	Some	All	Positional constraint
5	Adult-like	Some	Some	Combination of Types 3 and 4
6	Nonadult-like	All	All	Inventory constraint

Table 1. Types of productive phonological knowledge displayed by phonologically disordered children (from Gierut, 1985).

KNOWLEDGE TYPE	DESCRIPTION	EXAMPLE
1	A child displaying Type 1 knowledge of target [s] would produce this sound correctly in all word positions and for all morphemes; [s] would never be produced incorrectly.	[sʌn] "sun" [sup] "soup" [mɛsi] "messy" [mɪsɪŋ] "missing" [mɪs] "miss"
2	A child displaying Type 2 knowledge of target [s] would produce this sound correctly for all morphemes and positions. However, a phonological rule would apply to account for observed alternations between, for example, [s] and [t] in morpheme-final position.	[sʌn] "sun" [sup] "soup" [mɛsi] "messy" [aɪs] "ice" BUT: [mɪs] ~ [mɪt] "miss" [kɪs] ~ [kɪt] "kiss"
3	A child displaying Type 3 knowledge of target [s] would produce this sound correctly in all positions. However, certain morphemes that were presumably acquired early and acquired incorrectly (i.e., "fossilized") would always be produced in error.	[sʌn] "sun" [mɛsi] "messy" [mɪs] "miss" BUT: [nænə] "Santa" [vu] "juice"
4	A child displaying Type 4 knowledge of target [s] would produce this sound correctly for all morphemes in, for example, initial position. However, production of [s] would be incorrect for all morphemes in medial and final positions.	[sʌn] "sun" [sup] "soup" BUT: [mɛti] "messy" [mɪtɪŋ] "missing" [mɪt] "miss" [kɪt] "kiss"
5	A child displaying Type 5 knowledge of target [s] would produce this sound correctly in, for example, initial position. However, only some morphemes in this position would be produced correctly. All [s] morphemes in post-vocalic positions would be produced incorrectly.	[sʌn] "sun" [sup] "soup" BUT: [top] "soap" [tak] "sock" [mɛti] "messy" [kɪt] "kiss"
6	A child displaying Type 6 knowledge of target [s] would produce this sound incorrectly in all word positions and for all morphemes; [s] would never be produced correctly.	[tʌn] "sun" [tup] "soup" [mɪtɪŋ] "missing" [mɪt] "miss" [kɪt] "kiss"

Table 2. Description and examples of six types of productive phonological knowledge (from Gierut et al., in review).

target sound is classified according to the particular knowledge types described in Tables 1 and 2. The knowledge type that a child displays for a given target sound or class of sounds can be determined by answering the following questions regarding the nature of the child's lexical representation, the breadth of the distribution of sounds, and the use of phonological rules.

> Is the child's lexical representation of the target sound adult-like or nonadult-like? [2]
>
> If the child's lexical representation is adult-like,
>
> > does the correct representation extend to all word positions?
> >
> > does the correct representation extend to all morphemes?
> >
> > do phonological rules apply, thereby resulting in error productions?
>
> If the child's lexical representation is nonadult-like,
>
> > does the incorrect representation extend to all word positions?
> >
> > does the incorrect representation extend to all morphemes?
> >
> > do phonotactic constraints apply, thereby restricting production of the sound in certain positions (positional constraint) or excluding production of the sound from the inventory entirely (inventory constraint)?

Ranking productive phonological knowledge on a continuum

The first two steps in the assessment of phonological knowledge sample and establish what a child already knows about the target sound system; the third step establishes what a child has yet to learn. In this step, information about a child's phonological knowledge is organized in a systematic way. Specifically, a child's knowledge of target sounds is ranked on a continuum ranging from "most" to "least" phonological knowledge. A decision tree for ranking a child's knowledge on a continuum (Gierut, 1985) is presented in Figure 1. Notice that the decision tree takes into account the three primary factors used to establish productive phonological knowledge (i.e., lexical representation, distribution of sounds, and phonological rules). Also, the decision tree relates directly to the questions (above) used to determine productive phonological knowledge. Finally, notice that each of the six knowledge types is incorporated into the decision tree, as represented by the numbers shown on the diagonal.

Insert Figure 1 about here

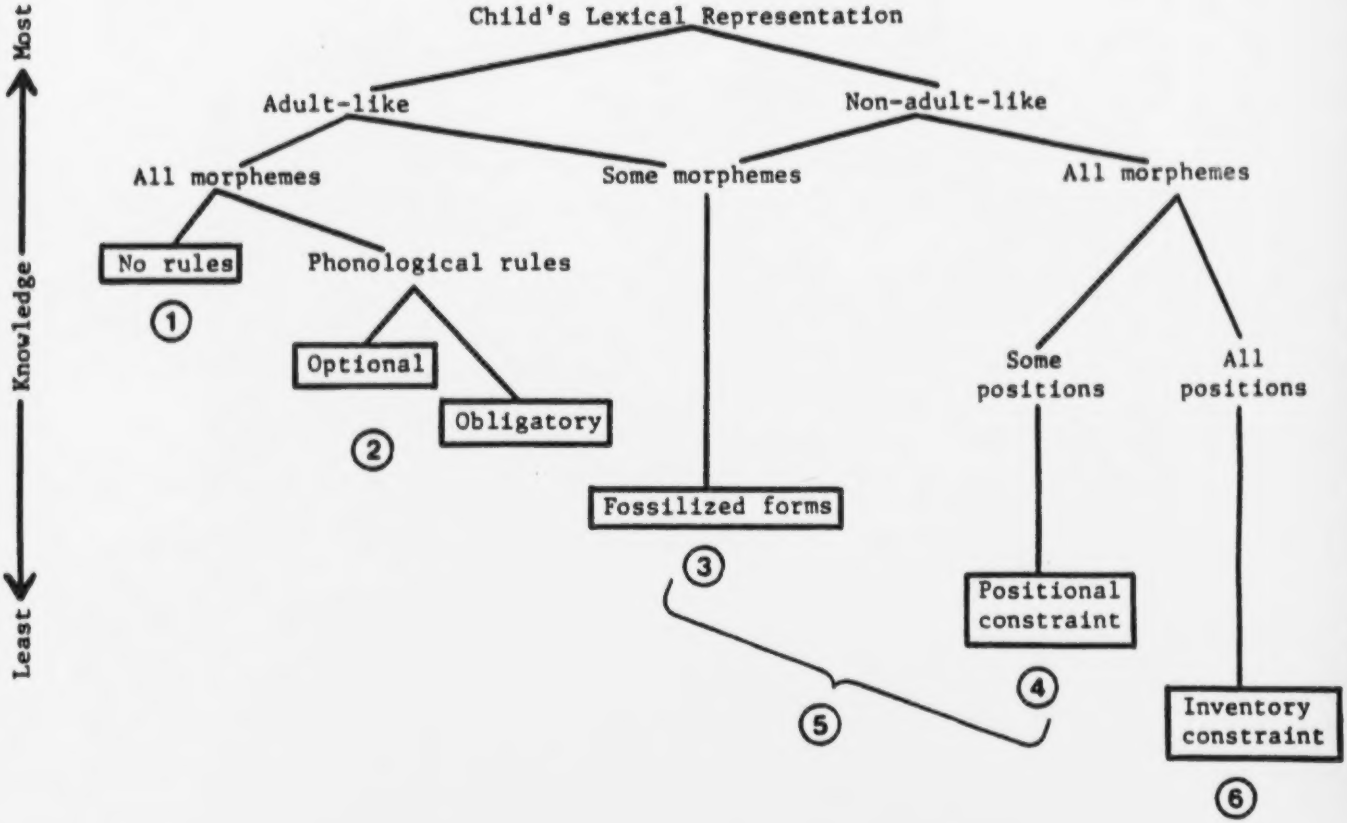

Most
Knowledge
Least
Child's Lexical Representation
Adult-like
Non-adult-like
All morphemes
Some morphemes
All morphemes
No rules
Phonological rules
①
Optional
Obligatory
②
Fossilized forms
③
Some positions
All positions
Positional constraint
④
⑤
Inventory constraint
⑥

Figure 1. Decision tree for ranking a child's productive phonological knowledge on a continuum ranging from "most" to "least" relative to the adult target sound system (from Gierut, 1985).

Using the decision tree to rank a child's knowledge of target sounds is relatively straightforward. Those sounds that a child produces in a manner similar to the adult, or Type 1 knowledge, are ranked at that end of the continuum labelled "most" knowledge. Those sounds that a child produces in a manner quite different from the adult, or Type 6 knowledge, are ranked at the opposite end of the continuum, "least" knowledge. Sounds classed as one of the intermediate knowledge types (Types 2 through 5) are ranked accordingly on the continuum.

Selecting treatment goals and sounds

Once a child's phonological knowledge of target sounds is determined and ranked on a continuum, treatment goals and sounds are selected directly from the continuum. Other auxiliary factors, such as age of acquisition of sounds, ease of production, or stimulability, are generally not considered in selecting treatment goals and sounds within this approach to assessment (see Gierut, 1986, for discussion). Treatment may procede sequentially beginning at one or the other end of the continuum.

Is starting treatment at one end of the continuum (e.g., "most" knowledge) more effective in interrupting a child's error pattern than starting treatment at the opposite end of the continuum (e.g., "least" knowledge)? Preliminary research (Gierut, 1985; Gierut et al., in review) has been conducted to address this question. The results of this research indicated that, when treatment began at the end of the continuum labelled "most" knowledge, children only learned treated sounds; other untreated sounds ranked lower on the knowledge continuum were not acquired. For example, when treatment began with sounds ranked high on the continuum at Type 2 knowledge, changes in production of these sounds were observed; however, production of other sounds ranked lower on the knowledge continuum, at Type 5 or Type 6 knowledge, did not change. Thus, when treatment begins with "most" knowledge, interruption of a child's error pattern is limited to the treated sounds; extensive reorganization of the overall sound system does not occur.

On the other hand, when treatment began with sounds ranked at the opposite end of the continuum, "least" knowledge, children learned treated sounds; in addition, other untreated sounds ranked higher on the knowledge continuum were also learned. For example, when treatment began with sounds ranked low on the continuum at Type 6 knowledge, changes occurred in production of these sounds; production of other sounds ranked higher on the knowledge continuum at Type 2 or Type 3 knowledge also improved. Thus, when treatment begins with "least" knowledge, interruption of a child's error pattern is extensive; widespread reorganization of the overall sound system occurs.

These research findings, while preliminary in nature, suggest that treatment should begin with sounds ranked at "least" knowledge in order to induce the greatest interruption in a child's error pattern. A treatment program beginning at the end of the continuum labelled "least" knowledge will be more effective than a program beginning at "most" knowledge.

Clinical Applications of Productive Phonological Knowledge

To illustrate how the assessment of productive phonological knowledge can be used in the clinic, the sound systems of two children have been analyzed. The two children were Annie, age 4 years, 6 months, and Clinton, age 4 years, 4 months. The children were considered to be functional misarticulators, producing several speech sounds in error from different sound classes with no known organic cause. Both children had normal hearing and auditory discrimination skills, were of normal intelligence, and were from monolingual English-speaking homes. The sound systems of these children, therefore, were particularly well-suited for the assessment of productive phonological knowledge. The assessment of productive phonological knowledge may be inappropriate in cases of children displaying only one or two sounds in error or errors of distortion. Moreover, the disordered sound systems of these children were not unusual or atypical, but rather, represent the types of errored sound systems that clinicians may frequently encounter in the clinic.

The analyses of each child's sound system follow the four steps in assessing productive phonological knowledge. Included in each child's analysis is a description of the type of phonological knowledge displayed for each class of target sounds, a continuum of productive phonological knowledge, and suitable target sounds for treatment. (It will be helpful for the reader to refer to Table 1 and Figure 1 throughout the discussion.)

Annie's Sound System

A representative sample of Annie's speech was obtained by having her tell stories about pictures in a book and by having her name common pictures and items according to the citation form procedure developed by Gierut (1985). The speech sample was phonetically transcribed and glossed and this sample served as the data base for assessing Annie's phonological knowledge.

Annie's phonetic inventory was relatively complete compared to the adult target and included the sounds:

m			n		ŋ	
p b			t d		k	ʔ
	f v	θ ð				
				tʃ dʒ		
w				j		h
			l	r		

Nasals. Annie accurately produced [m,n,ŋ] in all relevant word positions and for all target morphemes, indicative of Type 1 knowledge of these sounds.

Stops. Annie accurately produced the stops [p,b,d] relative to the adult in all word positions and for all morphemes. She demonstrated Type 1 phonological knowledge of these sounds.

Annie also produced the alveolar stop [t] in all word positions; however, for some morphemes, [t] alternated with either glottal stop or null in the word-final position. To illustrate, [t] was produced in alternation with a glottal stop in production of the morphemes "but" [bʌt] ~ [bʌʔ] and "got" [dat] ~ [daʔ]; [t] was produced in alternation with null in production of the morphemes "not" [nat] ~ [na] and "put" [pʊt] ~ [pʊ]. (The symbol "~" indicates an alternating production.) Notice, in these examples, that [t] was produced in final position, but that production of this sound varied with either a glottal stop or null. Annie was credited with phonological knowledge of [t], but to account for the observed variation, optional rules of word-final glottalization and word-final deletion applied. This is Type 2 phonological knowledge.

Annie's phonetic inventory included the velar stop [k] but this sound was only produced in the morpheme "ok." She never produced [k,g] in any other morphemes or in any word positions. [k,g] were always in error relative to the adult, being produced as [t,d] instead. Annie exhibited Type 6 phonological knowledge of [k,g].

Notice that Annie displayed several different types of phonological knowledge of target stops: Type 1 knowledge of [p,b], Type 2 knowledge of [t,d], and Type 6 knowledge of [k,g]. Of these sounds, Annie had relatively less knowledge of velar stops compared with bilabial or alveolar stops.

Fricatives. Annie produced the fricatives [f,v,θ,ð] correctly relative to the adult in all word positions and for all morphemes. She displayed Type 1 knowledge of these sounds. Annie never produced [s,z,ʃ] in any position or for any morphemes. These fricatives were always produced in error and were not in Annie's phonetic or phonemic inventories. Annie exhibited Type 6 knowledge of these sounds.

Affricates and Glides. Annie displayed Type 1 knowledge of [tʃ,dʒ] and [w,j,h], producing these sounds correctly in all word positions and for all morphemes.

Liquids. The liquid [l] was accurately produced in all word positions, but only for some target [l] morphemes. Other target [l] morphemes were consistently produced in error, indicative of Type 3 phonological knowledge. The liquid [r] was produced as a distortion. In this assessment framework, distortions indicate that a child's phonological knowledge is comparable to the adult target. Therefore, Annie was credited with Type 1 knowledge of [ɹ].

Annie's productive phonological knowledge of the sound system was then ranked on a continuum ranging from "most" (Type 1 knowledge) to "least" knowledge (Type 6 knowledge). The continuum of knowledge for this child is presented in Figure 2. Notice that all of the sounds for which Annie displayed Type 1 knowledge, i.e., [m,n,ŋ,p,b,d,f,v,θ,ð,tʃ,dʒ,w,j,h,r], were ranked at the top of the continuum. Ranked next, at Type 2 knowledge, was the target [t]; production of this sound was affected by optional phonological rules. The liquid [l] was ranked at Type 3 knowledge since Annie produced this sound for some, but not all, target morphemes. Annie did not display Type 4 or Type 5 knowledge of target sounds. Ranked lowest on the knowledge continuum were the sounds [k,g,s,z,ʃ] at Type 6 knowledge; these sounds were always in error relative to the adult.

Insert Figure 2 about here

In planning a remediation program for Annie, treatment goals and sounds can be systematically selected directly from the knowledge continuum. Three possible treatment goals were identified: (1) to eliminate the application of optional rules affecting production of [t], (2) to stabilize inconsistent productions of [l], and (3) to eliminate the inventory constraint affecting production of [k,g,s,z,ʃ]. Potential treatment sounds would, thus, include: (1) production of [t] in final position of words, (2) production of [l] in those morphemes consistently produced in error, and (3) production of [k,g,s,z,ʃ] in all word positions. While treatment can be initiated at any

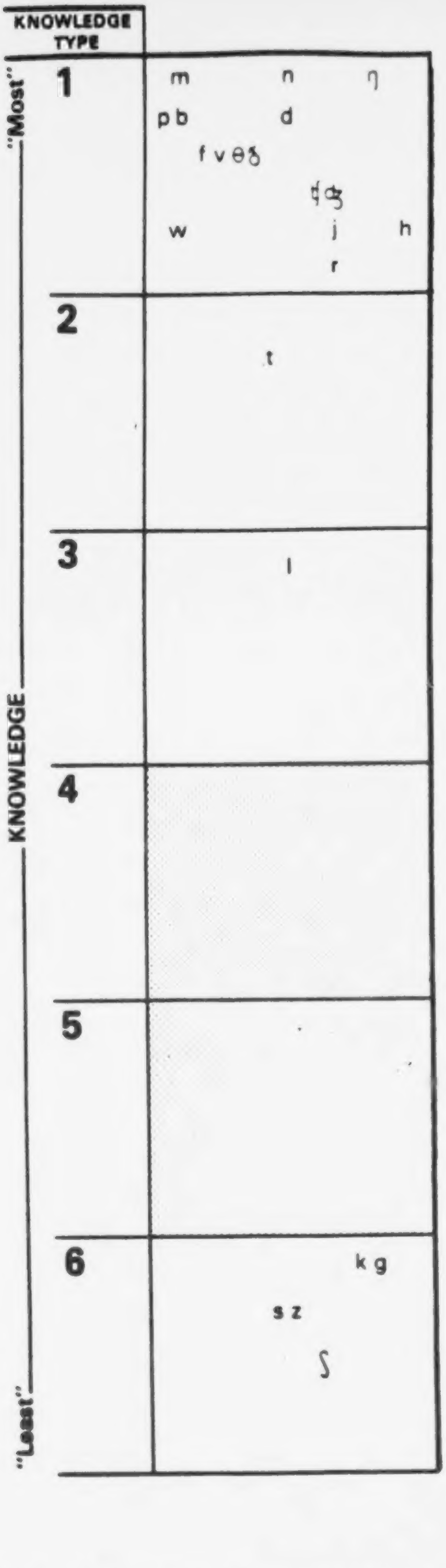


Figure 2. Continuum of productive phonological knowledge for Annie.

point along the knowledge continuum, it is recommended that treatment begin with those sounds ranked lowest on the knowledge continuum at "least" knowledge. Therefore, for Annie, treatment would begin with production of [k,g,s,z,ʃ]; these targets may be treated either singly or in combination. Starting treatment with these sounds should result in widespread reorganization of Annie's sound system.

Clinton's Sound System

Clinton's phonetic inventory included the sounds:

	m				n				ŋ	
p	b			t	d			k	g	ʔ
		f	v	s	z					
				t͜s	d͜z	tʃ	dʒ			
	w						j			h
					l		r			

Nasals. Clinton correctly produced the nasals [m,n,ŋ] for all morphemes. These sounds were never in error relative to the adult target. Clinton exhibited Type 1 knowledge of this class of sounds.

Stops. Clinton accurately produced the stops [p,b,t,d,k,g] in all word positions and for all target morphemes. Clinton, therefore, demonstrated Type 1 phonological knowledge of target stops.

Fricatives. The only fricatives produced correctly relative to the adult target were [s,z]. Clinton accurately used [s,z] in all word positions and for all target morphemes, indicative of Type 1 phonological knowledge.

The fricatives [θ,ð,ʃ], on the other hand, were never produced correctly in any position or for any target morphemes. These sounds were always in error relative to the adult and were not in Clinton's phonetic or phonemic inventories. Production of these sounds reflected Type 6 phonological knowledge.

Clinton's use of the fricatives [f,v] was somewhat more complicated. In word-initial position, [f] was used correctly for all target morphemes. [v], however, never occurred word-initially; [b] was used instead. Production of [v] was restricted from word-initial position indicating Type 4 phonological knowledge.

In post-vocalic positions, [f] was produced in alternation with [s], and [v] was produced in alternation with [z]. Clinton produced target [f] in two ways, either as [f] or as [s]; similarly, target [v] was produced either as [v] or as [z]. Production of either [f,v] or [s,z] was associated with phonetic context. Notice, in the examples that follow, that Clinton produced [f,v] in different ways depending on whether the sound was in the intervocalic or the final position of a morpheme.

[kɔwfɪn]	~	[kɔs]	"coughing"-"cough"
[jæfɪn]	~	[jæs]	"laughing"-"laugh"
[naɪfi]	~	[naɪs]	"knifey"-"knife"
[wevɪn]	~	[wez]	"waving"-"wave"
[sevɪn]	~	[sez]	"shaving"-"shave"
[dwaɪvɪn]	~	[dwaɪz]	"driving"-"drive"

Given these variable productions, what can we assess about Clinton's phonological knowledge? What did Clinton know about these particular morphemes and, more generally, what did he know about post-vocalic targets [f,v]? To answer these questions, consider, as an example, Clinton's production of the morphemes "cough" and "wave." The target fricatives [f,v] were produced (correctly) in these morphemes as [kɔf] and [wev], respectively; correct productions of [f,v] occurred in the intervocalic position when the present progressive suffix was added. The same exact morphemes were produced (incorrectly) as [kɔs] and [wez], respectively; incorrect productions of [f,v] occurred in the word-final position when the present progressive suffix was not added. Given these morphophonemic alternations, we will credit Clinton with phonological knowledge of post-vocalic targets [f,v] since he did produce these sounds correctly in the morphemes "cough" and "wave." However, since the variation between [f,s] and [v,z] was systematic and limited to a particular context, a phonological rule applied. This rule altered productions of [f,v] to [s,z] in word-final position. The rule was obligatory since it applied to all target [f,v] morphemes. Thus, Clinton exhibited Type 2 knowledge of [f,v].

Notice that Clinton exhibited several different types of phonological knowledge of target fricatives, i.e., Type 1 knowledge of [s,z], Type 2 knowledge of [f,v], Type 4 knowledge of word-initial [v], and Type 6 knowledge of [θ,ð,ʃ].

Affricates. Clinton exhibited Type 1 knowledge of the affricates [tʃ,dʒ]. These sounds were produced correctly relative to the adult target in initial position and as [t̯s,d̯z] in post-vocalic positions for all morphemes. (Remember that distortions indicate that a child's phonological knowledge is comparable to the adult.)

Glides. The sounds [w,j,h] were produced accurately in all word positions and for all target morphemes, reflecting Type 1 phonological knowledge.

Liquids. Clinton never produced targets [l,r] word-initially, but these sounds were produced accurately in post-vocalic positions. Type 4 knowledge accounted for the occurrence of liquids in some, but not all, word positions.

In addition to Type 4 knowledge of liquids, a phonological rule affected Clinton's production of post-vocalic [r]. Target [r] was produced either as [r] or as [w]. Production of [r] or [w] depended upon the phonetic context; [r] was produced in word-final position and [w] was produced in intervocalic position for the same exact morpheme. The examples that follow illustrate morphophonemic alternations between [r] and [w].

[dɪər]	~	[dɪːwi]	"deer"-"deery"
[tʃɛr]	~	[tʃɛwi]	"chair"-"chairy"
[stɔr]	~	[stɔwi]	"star"-"starry"

Clinton, therefore, was credited with phonological knowledge of post-vocalic target [r], but an obligatory rule altered production of [r] to [w] in the intervocalic position. The rule was obligatory since all post-vocalic target [r] morphemes were affected. This is Type 2 phonological knowledge.

Having assessed Clinton's productive phonological knowledge, a continuum of knowledge was developed and is displayed in Figure 3.

Insert Figure 3 about here

Notice that the sounds [m,n,ŋ,p,b,t,d,k,g,s,z,tʃ,dʒ,w,j,h] were all ranked at the top of the continuum at "most" phonological knowledge. Clinton exhibited Type 1 knowledge of these sounds, with accurate productions relative to the adult observed in all word positions and for all morphemes. Ranked next on the knowledge continuum were the sounds [f,v,r]; Clinton exhibited Type 2 knowledge of these sounds with phonological rules applying obligatorily. Next on the continuum, at Type 4 knowledge, were the sounds [v,l,r]. Clinton accurately produced these sounds in some, but not all, word positions. Finally, ranked lowest on the knowledge continuum were the sounds [θ,ð,ʃ]. These sounds were never produced in any position and were always in error relative to the adult. Clinton displayed Type 6 knowledge of these sounds. Notice that Clinton did not display Type 3 or Type 5 phonological knowledge.

Appropriate treatment goals and target sounds can be identified directly from Clinton's continuum of knowledge. Referring to Figure 3, three treatment goals were identified. One goal of treatment was to eliminate obligatory phonological rules (Type 2 knowledge) by targeting the sounds [f,v,r]. A second treatment goal was to eliminate positional constraints (Type 4 knowledge), targeting [v,l,r]. A third treatment goal was to eliminate inventory constraints; appropriate treatment targets in this case were [θ,ð,ʃ].

It is reasonable to implement a treatment program along any point on Clinton's continuum. As with Annie, however, the preferred remediation plan for Clinton would begin treatment at "least" knowledge, teaching the sounds [θ,ð,ʃ]. We would expect changes to occur not only in Clinton's production of these targets, but also, in other untreated sounds ranked higher on his knowledge continuum.

Conclusion

The evaluation of a child's productive phonological knowledge has direct applications for the three stage clinical process of assessment-diagnosis-treatment. Evaluating a child's phonological knowledge requires obtaining a representative speech sample, determining types of phonological knowledge, ranking phonological knowledge on a continuum, and selecting target sounds for treatment. As in other assessment frameworks, a child's pattern of error production is identified; then, through treatment, the pattern of error is interrupted. This framework for assessing phonological knowledge is different from other procedures, however, in that it examines a child's overall sound system, not just limited or isolated errors.

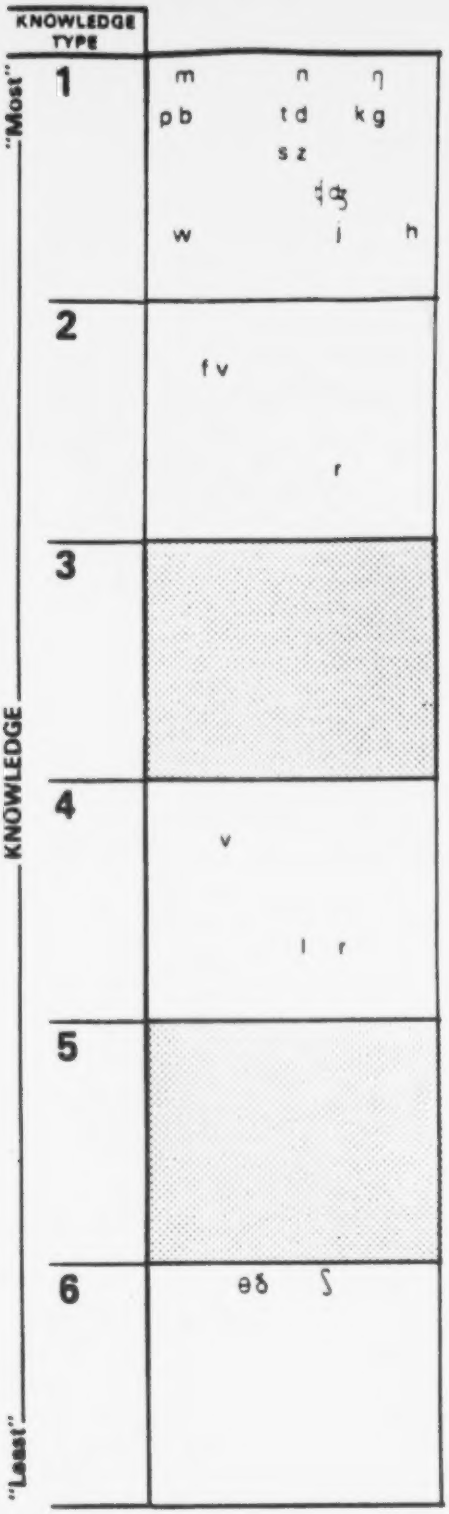

KNOWLEDGE TYPE
"Most"
KNOWLEDGE
"Least"
1
m n ŋ
p b t d k g
s z
tʃ dʒ
w j h
2
f v
r
3
4
v
l r
5
6
θ ð ʃ

Figure 3. Continuum of productive phonological knowledge for Clinton.

Also, this framework evaluates both a child's performance and competence of the sound system. Moreover, this framework has direct implications for treatment. Treatment goals and sounds are organized and selected in a principled manner based on a child's knowledge, and predictions about learning during treatment are generated. Thus, the framework for evaluating a child's productive phonological knowledge not only meets, but extends, the basic features of an adequate assessment procedure.

At present, however, several important questions about this assessment procedure remain unanswered (cf. Gierut, 1986). First, it is not known whether the assessment of productive phonological knowledge provides a more thorough evaluation or a more accurate diagnosis of speech sound disorders in children. Perhaps, all pattern analyses, while adopting different theoretical frameworks, are essentially equivalent in the assessment of speech sound disorders. Or, possibly, certain pattern analyses may be more appropriate for specific types of speech sound disorders. Second, it is not clear whether the assessment of productive phonological knowledge necessitates different models or strategies of treatment. Thus far, only established treatment procedures, e.g., minimal pair contrast treatment, have been implemented following assessments of phonological knowledge (cf. Gierut, 1985; Gierut et al., 1984; Gierut & Elbert, 1985; Gierut et al., in review). Third, it has yet to be determined whether the assessment of productive phonological knowledge is relevant to descriptions of developing sound systems. Possibly, the construct of productive phonological knowledge will provide new insights into normal phonological development, as well as phonological disorders. These, and other questions, remain open for future empirical study.

Endnotes

1 Generative phonological descriptions of a child's sound system rely on production data. It has also been suggested (cf. Barton, 1978) that speech perception or discrimination data may provide information about a child's phonological knowledge. There are, however, some inherent difficulties in using data from speech perception or discrimination to evaluate phonological knowledge. For example, Locke (1980a, 1980b) has reported that it is difficult to accurately and adequately assess a child's perceptual skills. Also, the role of perception or discrimination in learning sounds during treatment has not been clearly established (Williams & McReynolds, 1975; Winitz, 1975). Furthermore, recent evidence from primary languages, normal language development, speech disorders, and second language learning suggests that speech production and speech perception may be independent processes (Dinnsen, 1985; Straight, 1980).

2 In the literature on normal phonological development and phonological disorders, there has been some debate over the nature of children's lexical representations, either unique or adult-like (cf. Maxwell, 1984 for review). Within the generative approach to analysis, it has been demonstrated that phonologically disordered children may or may not evidence adult-like lexical representations (Dinnsen, 1984; Maxwell, 1981). For example, a child who omits final stops may evidence adult-like knowledge of these target sounds if he or she produces morphophonemically related forms with final stops, e.g., [da] "dog" ~ [dagi] "doggie" or [pɪ] "pig" ~ [pɪgi] "piggie." This child lexically represents morphemes in a manner comparable to the adult. Another child displaying a similar pattern of error, however, may not evidence adult-like knowledge of target stops if morphophonemically related forms do not alternate, e.g., [da] "dog" ~ [dai] "doggie" or [pɪ] "pig" ~ [pɪi] "piggie." This child lexically represents morphemes in a manner quite different from the adult.

References

Barton, D. (1978). *The role of perception in the acquisition of phonology*. Bloomington, IN: Indiana University Linguistics Club.

Bernthal, J., & Bankson, N. (1981). *Articulation disorders*. Englewood Cliffs, NJ: Prentice-Hall.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Dinnsen, D. A. (1984). Methods and empirical issues in analyzing functional misarticulations. In M. Elbert, D.A. Dinnsen, & Weismer, G. (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph, *22*, pp. 5-17). Rockville, MD: ASHA.

Dinnsen, D. A. (1985). A re-examination of phonological neutralization. *Journal of Linguistics*, *21*, 265-279.

Dinnsen, D. A., & Elbert, M. (1984). On the relationship between phonology and learning. In M. Elbert, D.A. Dinnsen, & Weismer, G. (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph, *22*, pp. 59-68). Rockville, MD: ASHA.

Elbert, M., Dinnsen, D. A., & Powell, T. W. (1984). On the prediction of phonologic generalization learning patterns. *Journal of Speech and Hearing Disorders*, *49*, 309-317.

Elbert, M., Dinnsen, D. A., & Weismer, G. (1984). *Phonological theory and the misarticulating child* (ASHA Monograph, *22*). Rockville, MD: ASHA.

Elbert, M., & Gierut, J. (1986). *Handbook of clinical phonology: Approaches to assessment and treatment*. San Diego: College-Hill Press.

Foss, D., & Hakes, D. (1978). *Psycholinguistics: An introduction to the psychology of language*. Englewood Cliffs, NJ: Prentice-Hall.

Gierut, J. A. (1985). *On the role of phonological knowledge and generalization learning in misarticulating children*. Doctoral dissertation, Indiana University, Bloomington. (Also distributed by the Indiana University Linguistics Club, Bloomington.)

Gierut, J. A. (1986). Generative phonology and clinical assessment frameworks: Empirical claims and differences. *Research on speech perception* (Progress Report 12, pp. 000-000). Bloominton, IN: Speech Research Laboratory, Department of Psychology.

Gierut, J. A., Dinnsen, D. A., & Elbert, M. (November, 1984). The implementation of a remediation program based on phonological knowledge. Presented at the American Speech-Language-Hearing Convention, San Francisco.

Gierut, J. A., & Elbert, M. (November, 1983). Phonological knowledge and the selection of training targets. Presented at the American Speech-Language-Hearing Convention, Cincinnati.

Gierut, J. A., & Elbert, M. (November, 1985). On the role of phonological knowledge in generalization learning. Presented at the American Speech-Language-Hearing Convention, Washington, D.C.

Gierut, J. A., Elbert, M., & Dinnsen, D. A. (in review). A functional analysis of phonological knowledge and generalization learning in misarticulating children. Manuscript submitted for publication to Journal of Speech and Hearing Research.

Ingram, D. (1981). Procedures for the phonological analysis of children's language. Baltimore: University Park Press.

Kenstowicz, M., & Kisseberth, C. (1979). Generative phonology: Description and theory. New York: Academic Press.

Locke, J. (1980a). The interference of speech perception on the phonologically disordered child. Part I: A rationale, some criteria, the conventional tests. Journal of Speech and Hearing Disorders, 45, 431-444.

Locke, J. (1980b). The interference of speech perception on the phonologically disordered child. Part II: Some clinically novel procedures, their use, some findings. Journal of Speech and Hearing Disorders, 45, 444-468.

Maxwell, E. M. (1981). A study of misarticulation from a linguistic perspective. Doctoral dissertation, Indiana University, Bloomington. (Also distributed by the Indiana University Linguistics Club, Bloomington.)

Maxwell, E. M. (1984). On determining underlying phonological representations of children: A critique of current theories. In M. Elbert, D.A. Dinnsen, & Weismer, G. (Eds.), Phonological theory and the misarticulating child (ASHA Monograph, 22, pp. 18-29). Rockville, MD: ASHA.

Maxwell, E. M., & Rockman, B. K. (1984). Procedures for linguistic analysis of misarticulated speech. In M. Elbert, D.A. Dinnsen, & Weismer, G. (Eds.), Phonological theory and the misarticulating child (ASHA Monograph, 22, pp. 69-84). Rockville, MD: ASHA.

McReynolds, L. V., & Engmann, D. (1975). Distinctive feature analysis of misarticulations. Baltimore: University Park Press.

Rockman, B. K. (1983). An experimental investigation of generalization and individual differences in phonological training. Unpublished doctoral dissertation, Indiana University, Bloomington.

Shriberg, L. D., & Kwiatkowski, J. (1980). Natural process analysis: A procedure for phonological analysis of continuous speech samples. New York: Wiley.

Straight, H. S. (1980). Auditory versus articulatory phonological processes and their development in children. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child phonology: Perception (Vol. 2, pp. 43-71). New York: Academic Press.

Turton, C. J. (1973). Diagnostic implications of articulation testing. In W. D. Wolfe & D. J. Goulding (Eds.), *Articulation and learning* (pp. 195-232). Springfield, IL: Charles C. Thomas.

Williams, G., & McReynolds, L. V. (1975). The relationship between discrimination and articulation training in children with misarticulations. *Journal of Speech and Hearing Research*, *18*, 401-412.

Winitz, H. (1975). *From syllable to conversation*. Baltimore: University Park Press.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Generative Phonology and Error Pattern Analyses: Empirical Claims and Differences*

Judith A. Gierut

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana

---

*This research was supported, in part, by NIH Training Grant NS-07134 to Indiana University in Bloomington. I would like to thank Daniel Dinnsen for his suggestions on an earlier version of this manuscript and Kathy Hoyt for her assistance with the data analyses.

Abstract

Clinical assessment is a three component process involving characterization of the problem, selection of treatment targets, and projection about learning. In the assessment of functional speech sound disorders, several different approaches have been used, the most recent being generative phonological analysis. This paper examines the claims of generative phonology underscoring the differences between various analysis procedures. The generative approach is shown to be empirically different by characterizing speech sound disorders in terms of a child's productive phonological knowledge, by selecting treatment sounds based on a continuum of productive phonological knowledge, and by predicting system-wide changes in the sound system. The clinical importance of these differences is discussed.

Generative Phonology and Error Pattern Analyses:

Empirical Claims and Differences

Central to the clinical management of functional speech disorders is an accurate and thorough assessment of the problem. To a large extent, the relative success or failure of an intervention program depends on assessment; treatment can only be as effective as the evaluation is accurate. In assessment, a clinician establishes what a speaker knows about the target language and identifies what remains to be learned during intervention. A clinician also recommends appropriate treatment goals and plans an intervention program that promotes maximal learning. Assessment can, therefore, be considered a three component process involving: (1) characterization of the problem, (2) selection of treatment goals, and (3) projection about learning.

In the clinical management of functional speech sound disorders in children, in particular, there have been several frameworks, or models of assessment that have been adopted. These include various error pattern analyses such as place-voice-manner analysis (Turton, 1973; Weber, 1970), distinctive feature analysis (McReynolds & Engmann, 1975; Singh, 1976), and phonological process analysis (Hodson, 1980; Ingram, 1981; Shriberg & Kwiatkowski, 1980; Weiner, 1979). More recently, standard generative phonology has been introduced as an alternate framework for error pattern analysis (cf. Elbert, Dinnsen, & Weismer).[1] The focus of this paper is to compare the empirical claims of generative phonology with those of the other non-generative pattern analyses.

Non-generative pattern analyses

Place-voice-manner analysis, distinctive feature analysis, and phonological process analysis each corresponds, historically, to a shift in the clinical approach to analyzing sound errors. It is not clear, however, whether the different frameworks uniquely distinguish themselves in terms of the three components in the assessment process.

For instance, in characterization, place-voice-manner analysis describes the pattern underlying a child's error productions using a three-way articulatory phonetics classification system. Sound errors are not viewed as random and isolated, but are considered systematic and consistent. Likewise, distinctive feature analysis emphasizes patterns of production. This framework applies concepts from linguistic theory to describe the regularity of children's errors; it also borrows principles from learning theory in the treatment of feature-related sounds. Phonological process analysis similarly identifies error patterns using elements of linguistic theory. Within this framework, error patterns are characterized as phonological rules that describe systematic correspondences between a child's production and an adult target. The focus of treatment, then, is on the elimination of phonological processes, not on the acquisition of specific sounds or features as in other pattern analyses. While each model of assessment offers a variation in the approach to characterizing sound errors, these differences may only represent terminological shifts (see Elbert, 1985; Locke, 1983; McReynolds & Elbert, 1981a; Shelton & McReynolds, 1979; Shriberg & Kwiatkowski, 1982a for discussion).

Also, the selection of target sounds for treatment is essentially equivalent across the different analyses. Elbert and Gierut (1986), for example, analyzed the speech of a phonologically disordered child, A.B., within the framework of three pattern analyses, place-voice-manner, distinctive feature, and phonological process analysis. A.B.'s pattern of error was characterized as involving manner of production (place-voice-manner analysis), continuant and strident features (distinctive feature analysis), and the process of stopping (phonological process analysis). In each case, however, the recommended goal of treatment was production of fricatives. The approach to characterizing A.B.'s error pattern was different across the various frameworks, but the preferred goal of treatment was identical.

Finally, predictions about learning are comparable across different models of assessment. Intervention programs are typically structured to interrupt the pattern underlying a child's productions by promoting a maximum amount of learning (i.e., generalization) following a minimum amount of treatment. Various types of generalization learning have been consistently observed both within and across different pattern analyses, e.g., generalization to untreated word positions (Elbert & McReynolds, 1975, 1978; McReynolds, 1972), to untreated sounds within a class (Elbert, Shelton, & Arndt, 1967; Costello & Onstine, 1976; Weiner, 1981), to linguistic units of increasing complexity (McLean, 1970; Wright, Shelton, Arndt, 1969), and to new listeners and settings (Costello & Bosler, 1976; Olswang & Bain, 1985). Generalization learning appears to be relatively independent of the analysis procedure that is used. Regardless of the assessment framework, a clinician can project the kind of generalization that will occur following treatment, as well as how generalization will affect the pattern of error (see Elbert & Gierut, 1986, for discussion).

Generative pattern analysis

As an initial overview, generative phonology appears to be similar to existing analyses in that it describes and interrupts a child's pattern of production using principles of theoretical linguistics and learning theory. Generative phonology seems to be different in that it describes a child's sound system independent of the adult target. This approach not only identifies surface-level errors, but also describes a child's underlying knowledge of the sound system. Generative phonology recommends target sounds, structures treatment programs, and predicts learning on the basis of underlying or productive phonological knowledge.

The generative approach to analysis appears to be unique in comparison to other models of assessment. It is necessary, however, to empirically determine whether generative phonology is, in fact, different from existing pattern analyses. Does the generative approach to assessment add to the characterization and treatment of functional speech sound disorders? Or, does the generative approach to assessment represent no more than just a relabeling of surface error patterns? At the very least, the generative approach should provide information comparable to existing analyses about the nature of speech sound disorders. At best, the generative approach should add to and extend the current state of information about this disorder. Thus, the purpose of this paper is to evaluate the generative approach to assessment with respect to other types of pattern analyses. The empirical claims of the generative approach are discussed relative to differences in the three components in the assessment process: characterization of the problem, selection of treatment sounds, and predictions about learning during treatment.

Characterization of the Problem

The initial step in any model of assessment involves identifying, describing, and characterizing a child's pattern of error. The procedures for the characterization of error patterns within a generative framework have been described in detail elsewhere (Dinnsen, 1984; Elbert & Gierut, 1986; Gierut, 1985b, in press; Gierut, Elbert, & Dinnsen, in review; Maxwell & Rockman, 1984) and are only summarized herein. Briefly, generative phonology characterizes a child's production pattern by evaluating both the competence and performance (cf. Foss & Hakes, 1978) of the child's sound system. The assessment of performance provides a surface level evaluation of a child's sound system by identifying: (1) those sounds that are produced regardless if correct (i.e., phonetic inventory), (2) those sounds that are used contrastively to signal meaning differences (i.e., phonemic inventory), and (3) the distribution of sounds by word position and morphemes. The evaluation of performance establishes the structure of sounds in a child's sound system. This aspect of generative analysis is similar to other pattern analyses. The assessment of competence, however, is particular to generative analysis. Here, the underlying use and function of sounds is identified. The assessment of competence provides a characterization of a child's lexical, or underlying representation of morphemes, in addition to the use of phonological rules. Lexical representations of morphemes refer to the idiosyncratic, unpredictable properties of productive language, learned and stored in a speaker's lexicon; phonological rules describe predictable variations in production and associate sound to meaning (see Dinnsen, 1984; Gierut, in press, for more detail).

Thus, the assessment of performance provides information about the inventory and distribution of sounds; the assessment of competence provides information about unpredictable (i.e., lexical representations of morphemes) and predictable (i.e., phonological rules) aspects of productive language. When taken together, these components define a child's productive phonological knowledge (Dinnsen & Elbert, 1984) of the sound system.

Certain types of data are required to evaluate productive phonological knowledge.[2] As in other pattern analyses, the generative framework examines production of all target English sounds in each of three word positions (initial, medial, and final) by sampling both connected speech and spontaneously produced citation forms. The generative approach also samples each sound in each word position in more than one morpheme and further samples each morpheme more than one time. Unique to the data base, however, is the elicitation of potential morphophonemic alternations (e.g., "pig" - "piggy," "kiss" - "kissing"). Morphophonemic alternations illustrate systematic changes in production of a morpheme when placed in different phonological contexts and thereby provide evidence for how sounds function in that context. Morphophonemic alternations also provide information about a child's lexical representation of morphemes and the use of phonological rules.

The evaluation of productive phonological knowledge also involves classifying a child's use of each target sound as one of six different knowledge types, as in Tables 1 and 2 (Gierut, 1985b, in press; Gierut, Elbert, & Dinnsen, 1985; Gierut et al., in review). The six knowledge types derive from the three factors above associated with competence and performance: (1) the nature of a child's lexical representation of morphemes, either adult-like or nonadult-like; (2) the breadth of the distribution of sounds, either to some or all word positions, or to some or all morphemes; and (3) a child's use of phonological rules.

Insert Tables 1 and 2 about here

The knowledge types describe the full range of consistency and variability observed in a child's production of sounds. Notice that sounds always produced correctly relative to the (adult) target are characterized as type 1 knowledge; sounds always produced incorrectly are characterized as type 6 knowledge. Other systematic variations in sound production are characterized as knowledge types 2 through 5. This systematic characterization of variable productions based upon factors associated with competence and performance is another unique component of generative analysis. Other pattern analyses may also identify variation in sound production through measures such as sound production tasks (Elbert et al., 1967) or deep tests (McDonald, 1964). In these cases, however, the _source_ of inconsistent productions (e.g., nonadult-like lexical representation of some morphemes) is not identified. Also, distinctions among different _types_ of inconsistency (e.g., by morpheme, or word position, or across repetitions) are not noted. Finally, the _systematicity_ of variable productions is not described with respect to particular phonological contexts or morphemes. For example, Camarata and Gandour (1984) observed systematic changes in a child's inconsistent production of alveolar and velar stops, but only when these sounds were examined in phonologically relevant contexts. Similarly, Fey and Stalker (in press) identified systematic changes in a child's inconsistent production of sounds in specific morphemes, but only after considering morphophonemic alternations.

To summarize, generative analysis is different from other pattern analyses in several ways. First, a child's pattern of error is assessed by examining competence and performance of the sound system. Second, the evaluation of competence and performance requires that not only the production of sounds at a phonetic level, but also the function of sounds at a phonological level, is evaluated. Third, the data base for this analysis includes elicitation of potential morphophonemic alternations. Finally, a child's production of each target sound is characterized as a specific type of productive phonological knowledge relative to the adult target. Both the characterization of a child's production pattern and the procedures of generative analysis are different from other analyses. Further differences emerge when the selection of target sounds and predictions about learning are considered, as will be seen below.

Selection of Treatment Sounds

The second component in the assessment process involves selecting target sounds for treatment. Many different selection factors are typically taken into account in pattern analyses, e.g., age of acquisition of target sounds, frequency of errors, stimulability, overall intelligibility. Any combination of these, or other factors may enter into target sound selection. However, to date, there are no available guidelines for the optimal selection of treatment sounds. It has not yet been determined whether selection factors are of equal weight in pattern analyses; perhaps, some factors are more important than others in cases of particular children or specific error types.

KNOWLEDGE TYPES	LEXICAL REPRESENTATION	BREADTH OF DISTRIBUTION		PHONOLOGICAL RULE ACCOUNT
		Positions	Morphemes	
1	Adult-like	All	All	None
2	Adult-like	All	All	Optional or obligatory rules
3	Adult-like	All	Some	Fossilized forms
4	Adult-like	Some	All	Positional constraint
5	Adult-like	Some	Some	Combination of Types 3 and 4
6	Nonadult-like	All	All	Inventory constraint

Table 1. Types of productive phonological knowledge displayed by phonologically disordered children (from Gierut, 1985b).

KNOWLEDGE TYPE	DESCRIPTION	EXAMPLE
1	A child displaying Type 1 knowledge of target [s] would produce this sound correctly in all word positions and for all morphemes; [s] would never be produced incorrectly.	[sʌn] "sun" [sup] "soup" [mɛsi] "messy" [mɪsɪŋ] "missing" [mɪs] "miss"
2	A child displaying Type 2 knowledge of target [s] would produce this sound correctly for all morphemes and positions. However, a phonological rule would apply to account for observed alternations between, for example, [s] and [t] in morpheme-final position.	[sʌn] "sun" [sup] "soup" [mɛsi] "messy" [aɪs] "ice" BUT: [mɪs] ~ [mɪt] "miss" [kɪs] ~ [kɪt] "kiss"
3	A child displaying Type 3 knowledge of target [s] would produce this sound correctly in all positions. However, certain morphemes that were presumably acquired early and acquired incorrectly (i.e., "fossilized") would always be produced in error.	[sʌn] "sun" [mɛsi] "messy" [mɪs] "miss" BUT: [nænə] "Santa" [vu] "juice"
4	A child displaying Type 4 knowledge of target [s] would produce this sound correctly for all morphemes in, for example, initial position. However, production of [s] would be incorrect for all morphemes in medial and final positions.	[sʌn] "sun" [sup] "soup" BUT: [mɛti] "messy" [mɪtɪŋ] "missing" [mɪt] "miss" [kɪt] "kiss"
5	A child displaying Type 5 knowledge of target [s] would produce this sound correctly in, for example, initial position. However, only some morphemes in this position would be produced correctly. All [s] morphemes in post-vocalic positions would be produced incorrectly.	[sʌn] "sun" [sup] "soup" BUT: [top] "soap" [tak] "sock" [mɛti] "messy" [kɪt] "kiss"
6	A child displaying Type 6 knowledge of target [s] would produce this sound incorrectly in all word positions and for all morphemes; [s] would never be produced correctly.	[tʌn] "sun" [tup] "soup" [mɪtɪŋ] "missing" [mɪt] "miss" [kɪt] "kiss"

Table 2. Description and examples of six types of productive phonological knowledge (from Gierut et al., in review).

Also, within models of assessment, there is little agreement on the selection metric that should be used. Within the phonological process approach, for example, it has been recommended that the treatment of processes should be based on the frequency of process use (Hodson, 1980; Ingram, 1981). It has also been suggested that the selection of processes to be treated should be based on developmental sequence; i.e. those phonological processes first eliminated by normally developing children should be treated first (Edwards & Bernhardt, 1973; Ingram, 1981; Shriberg & Kwiatkowski, 1980; Weiner, 1979). Another recommendation has been that phonological processes which apply optionally, or processes which are used inconsistently, should be targeted for treatment initially (Edwards & Bernhardt, 1973).

Selecting treatment sounds becomes a difficult clinical decision, then, given the lack of criteria for sound selection and the range of factors considered within particular pattern analyses. Consequently, the selection of target sounds may be based to a large degree on a clinician's subjective judgment.

Alternate criteria for selecting treatment targets are used in the generative framework. Generative phonology does not necessarily exclude such factors as age of acquisition, stimulability, or frequency of errors; however, these considerations are auxiliary. Because generative phonology was not designed specifically for use in speech-language pathology as a means of assessing children's sound errors, typically used selection factors are less relevant. In the generative framework, treatment sounds are selected on the basis of a child's productive phonological knowledge and a hierarchical relationship among sounds (Gierut, 1985b, in press; Gierut et al., 1985). Specifically, a child's productive phonological knowledge is arranged on a continuum ranging from "most" to "least" knowledge of target sounds. Sounds classed as type 1 knowledge are ranked at the end of the continuum labeled "most" knowledge; sounds classed as type 6 knowledge are ranked at the opposite end of the continuum, "least" knowledge. Other sounds classified as one of the intermediate knowledge types 2 through 5 are ranked accordingly. Selection of target sounds for treatment then derives directly from this continuum. Targets are chosen from either end of the continuum, with treatment proceeding sequentially from "most" to "least" knowledge or from "least" to "most" knowledge. It should be noted, however, that the entire knowledge continuum does need not be treated in order for a child's error pattern to be interrupted. Experimental data (Gierut, 1985b) on this point suggests that treatment can be limited to a few target sounds at either end of the continuum. Thus, the selection of treatment sounds within generative phonology is based upon a hierarchical relationship that emerges among the different types of productive phonological knowledge that a child displays about target sounds.

The selection of treatment targets within the generative framework is, in principle, unique. It is fair to ask, however, whether the knowledge continuum corresponds in any way to selection factors considered in other pattern analyses. For example, does the knowledge continuum reflect a developmental sequence of sound acquisition, or frequency of errors? If this were the case, the knowledge continuum would only provide an alternate but comparable means of selecting treatment sounds. If, on the other hand, a correspondence were not observed among selection factors, then the knowledge continuum would represent a novel method of choosing treatment sounds.

To evaluate this potential equivalence, the knowledge continuum is compared to selection factors used in other pattern analyses. Continua of knowledge from six different children are presented as examples. These continua were selected as representative of those reported elsewhere (Gierut, 1985a, 1985b). The continua were all developed in the same way, using both spontaneous connected speech and citation form samples, and the procedures described for classifying and ranking productive phonological knowledge (Gierut, 1985b, in press; Gierut et al., 1985). Mean point-to-point inter- and intrajudge transcription reliability reported on these samples (Gierut, 1985b; Gierut et al., in review) was 98% and 93% agreement, respectively.

Continuum of knowledge as an index of acquisition sequence

Age of acquisition of sounds is one factor often considered in selecting treatment targets. Sounds that should be, but are not yet mastered (according to developmental norms) by a child of a given age may be chosen for treatment. Perhaps, the knowledge continuum corresponds to age of acquisition, with sounds mastered early ranked at "most" knowledge and sounds mastered later ranked at "least" knowledge.

The knowledge continua for two children, ages 4 years, 6 months (Child 1) and 4 years, 4 months (Child 2) are presented in Figure 1. Comparing these data to developmental norms (Prather, Hedrick, & Kern, 1975), it can be seen that the knowledge continua do not directly correspond to age of acquisition. First, sounds typically mastered early were ranked at "least" knowledge. Child 1, for example, demonstrated "least" knowledge of [k,g], sounds generally acquired by age 3. Second, sounds mastered late were ranked at "most" knowledge. Both Child 1 and 2 demonstrated relatively "most" knowledge of targets [v,tʃ,dʒ,r], mastered beyond age 4. Third, sounds mastered at approximately the same age were ranked at opposite ends of the knowledge continuum. The sounds [b,g,f] are all typically acquired by age 3; for Child 1, [b,f] were ranked at "most" knowledge, whereas [g] was ranked at "least" knowledge. Similarly, [ʃ,tʃ,dʒ] are all mastered at approximately the same age, but for Child 2, [tʃ,dʒ] were ranked at "most" knowledge and [ʃ] was ranked at "least" knowledge. Finally, as might be expected, some early-mastered sounds (e.g., nasals) were ranked at "most" knowledge and other late-mastered sounds (e.g., [ð,ʃ]) were ranked at "least" knowledge for both children. Thus, the data from these two children illustrate that the knowledge continuum does not directly correspond to age of sound acquisition. A knowledge continuum is not, therefore, a notational variant of an age of acquisition hierarchy.

Insert Figure 1 about here

Continuum of knowledge as an index of quantitative measures of sound accuracy

Quantitative measures, such as baseline scores or percentages of accurate production, are other factors considered in sound selection. Sounds frequently and consistently in error are sometimes chosen for treatment, since remediation of these sounds significantly interrupts an error pattern (cf. Bernthal & Bankson, 1981). On the other hand, sounds infrequently and inconsistently in error have also been selected for treatment (cf. Bernthal & Bankson, 1981). The intent, here, is to have a child achieve initial success before advancing to more frequent errors and, perhaps, more difficult sounds. It may be tempting to think of the knowledge continuum as corresponding to quantitative measures of accurate sound production. Sounds ranked at "most" knowledge may be indicative of a high percentage of accuracy, whereas sounds

"Most" — KNOWLEDGE — "Least"

KNOWLEDGE TYPE	CHILD 1	CHILD 2
1	m n ŋ p b d f v θ ð tʃ dʒ w j h r	m n ŋ p b t d k g s z tʃ dʒ w j h
2	t	f v r
3	l	▨
4		v l r
5		▨
6	k g s z ʃ	θ ð ʃ

Figure 1. The knowledge continua for Children 1 and 2.

ranked at "least" knowledge may reflect a low percentage of accuracy. To evaluate this hypothesis, the knowledge continuum is compared to two different quantitative measures, i.e. baseline scores and percentage of consonants correct (PCC) (Shriberg & Kwiatkowski, 1982b).

Baseline scores. Baseline scores have been used as one measure of a child's productive phonological knowledge. Elbert, Dinnsen, and Powell (1984), for instance, equated baseline performance with productive phonological knowledge to distinguish between "phonologically known" versus "phonologically unknown" sounds. In this study, generative analyses of children's sound systems were not developed; rather, an equivalence relationship was established between baseline scores and productive phonological knowledge.[3] Given that generative analyses are qualitative and descriptive, and rely on specific types of data, i.e., morphophonemic alternations, it is highly likely that baseline scores do not, in all cases, accurately reflect productive phonological knowledge as ranked on a continuum.

Table 3 presents the ranking of sounds on a knowledge continuum for three children, in addition to the baseline scores obtained for these same sounds prior to treatment. The relative ranking of sounds on the knowledge continua was established independent of baseline scores. As described above, the knowledge continua were developed on the basis of descriptive generative analyses; baseline scores represent the percentage of accurate sound production on a probe measure consisting of spontaneously produced citation forms (Gierut, 1985b).

Insert Table 3 about here

Notice, first, that baseline scores corresponded to the relative ranking of sounds on the knowledge continuum in the case of Child 3. Baseline data accurately reflected Child 3's productive phonological knowledge. The sound [t], ranked at "most" knowledge, was also produced with the greatest accuracy (i.e., 60%); likewise, [s], ranked at "least" knowledge, was produced with the least accuracy (i.e., 0%). The hierarchical relationship among sounds on the continuum was reflected in Child 3's performance in baseline. This observation is consistent with Elbert et al. (1984).

Baseline scores, however, did not correspond to qualitative claims about productive phonological knowledge in the case of Child 4. The ranking of sounds on this child's continuum, from "most" to "least" knowledge, was [ʃ] -> [dʒ] -> [v]. Baseline performance indicated that production of [dʒ] (i.e., 85%) was more accurate than production of [ʃ] (i.e., 60%) which, in turn, was more accurate than production of [v] (i.e., 0%). Here, baseline scores and productive phonological knowledge were not equivalent. This mismatch relates to the type of knowledge that Child 4 displayed for these target sounds. The generative analysis credited Child 4 with type 3 knowledge of [ʃ], or adult-like lexical representations in all word positions but only for *some* morphemes. Here, the accuracy of [ʃ] production was associated with *specific* morphemes. Items sampled on the baseline measure were most likely those target [ʃ] morphemes that Child 4 represented in a nonadult-like manner; hence, the lower baseline score. Methodologically, it would be impossible, of course, to identify all the morphemes that Child 4 represented in an adult-like versus a nonadult-like way for purposes of constructing and

PRODUCTIVE PHONOLOGICAL KNOWLEDGE

"Most" --> "Least"

CHILD 3 (4;3 years)	[t] 60% (Type 2)	[ʃ] 45% (Type 3)	[s] 0% (Type 6)
CHILD 4 (4;3 years)	[ʃ] 60% (Type 3)	[dʒ] 90% (Type 4)	[v] 30% (Type 4)
CHILD 5 (4;6 years)	[k] 0% (Type 6)	[s] 0% (Type 6)	[ʃ] 0% (Type 6)

Table 3. Sounds selected for treatment for three children as ranked on a continuum ranging from "most" to "least" knowledge. Baseline scores obtained prior to treatment for these same sounds are also reported as percentages of accurate production. The type of productive phonological knowledge each child displayed for these targets is noted in parentheses.

obtaining a representative baseline measure. The accuracy of [dʒ,v], on the other hand, was associated with word position. Child 4 demonstrated type 4 knowledge of these sounds, or adult-like lexical representations in some positions for *all* morphemes. In this case, it was possible to obtain a representative baseline measure by sampling specific word positions. For Child 4, then, a different ranking of sounds on the knowledge continuum would have resulted if baseline data, rather than generative descriptions, had been used to establish productive phonological knowledge.

Also, baseline data did not adequately identify differences in productive phonological knowledge in the case of Child 5. The knowledge continuum differentiated among Child 5's productive phonological knowledge of [k,s,ʃ]; however, baseline scores for these three sounds were identical at 0% accuracy. This mismatch between baseline scores and the knowledge continuum related to the nature of the child's lexical representations. Specifically, Child 5 represented target [k] as /t/; the child's lexical representation and the target sound were both from the same class. Child 5 also represented target [s] as /t/; the lexical representation and target sound were from different classes. Child 5 lexically represented [ʃ] as either /tʃ/ or /t/ depending on the morpheme and word position; e.g., "shoe" /tʃu/, "wash" /wɔt/. Notice that, for all three targets, Child 5 displayed nonadult-like lexical representations, but was credited with subtle differences in knowledge based on how these sounds were represented; i.e., representations from the same or different sound class as the target, or representations from one or more than one sound class. For Child 5, then, baseline data were not as sensitive as the continuum of knowledge in identifying differences in productive phonological knowledge of sounds.

One other instance where baseline scores are not equivalent to productive phonological knowledge is that of a child displaying type 2 knowledge, or the use of phonological rules (Gierut, 1986). In this case, the generative analysis credits a child with adult-like knowledge of target sounds, or relatively "most" knowledge. Baseline scores, however, reflect the inaccurate production of these target sounds as a result of the application of a phonological rule. The mismatch results because baseline measures only sample how sounds are produced, i.e., the correct/incorrect production of sounds; whereas, generative analysis samples how sounds are represented, distributed, and function in a child's sound system.

These comparisons illustrate that baseline scores do not accurately or adequately establish a child's productive phonological knowledge in all cases.[4] Furthermore, these data suggest that treatment sounds selected from a knowledge continuum would not necessarily be the same as those selected using baseline scores.

Percentage of consonants correct. Another quantitative measure that has been described in the literature (Shriberg & Kwiatkowski, 1982b) involves calculating a percentage of consonants correct (PCC). The PCC metric provides a measure of the accuracy of a child's production of target English sounds in each of three word positions based on a connected speech sample. As was suggested with baseline scores, it may be thought that the continuum of knowledge is equivalent to PCC values.

Table 4 presents the knowledge continuum and PCC values for Child 6, age 3 years, 7 months. The knowledge continuum and PCC values were established independently. As described, the knowledge continuum was developed from an extended connected speech sample and spontaneously produced citation forms that included morphophonemic alternations. PCC values were calculated from a

3-minute portion of the same connected speech sample (Appendix A) in accord with PCC procedures (Shriberg & Kwiatkowski, 1982b).

Insert Table 4 about here

First, notice that for both the knowledge continuum and PCC metric, four divisions or groupings of sounds emerged. The knowledge continuum grouped sounds according to type of productive phonological knowledge, while the PCC metric grouped sounds on a scale of severity of involvement. Certain sounds within these divisions were comparable across the knowledge continuum and PCC metric. For example, the sounds [m,n,ŋ,k,s,w,h] were classed as type 1 ("most") knowledge and, similarly, as mild in severity. Also, [f,v,ð,ʃ,tʃ,ʤ,l] were classed as type 6 ("least") knowledge, or severely involved. This overlap was noted for the categories "most" knowledge/"mild" involvement and "least" knowledge/"severe" involvement. The overlap, however, was not all-inclusive. This observation is not surprising given that knowledge types 1 and 6 define sounds that are consistently produced correctly or incorrectly, relative to the adult.

Second, sounds ranked on the knowledge continuum did not always show a one-to-one correspondence with sounds classed together on the PCC metric. In these instances, the generative approach credited the child with more productive phonological knowledge than PCC scores indicated. As an example, [b] was ranked as type 1 knowledge on the continuum; Child 6 consistently produced and used this sound correctly relative to the adult. PCC values, however, indicated that [b] was moderately to severely involved. A similar case resulted with the targets [z,j]. These sounds were ranked as type 1 knowledge on the continuum, but were classified as severely involved on the PCC metric. This discrepancy is likely related to the fact that all sound errors are not viewed as equivalent in the generative framework. For example, errors that result from the application of phonetic implementation rules (e.g., distortions) are considered less severe than errors that result from the application of phonological rules (e.g., omissions). Similarly, errors that are the result of the application of phonological rules are less severe than errors that relate to the nature of a child's lexical representation. The generative framework makes distinctions among errors at the phonological level versus those at the phonetic level of representation. The PCC metric, on the other hand, views all errors, regardless of the source and type, as equivalent. This discrepancy may also be further related to methodological differences involving the nature of the sample and the kinds of data considered in each analysis.

In these examples, the generative approach attributed more knowledge to the child than was determined by the PCC metric. The reverse situation, however, was not observed. That is, there were no cases where the generative framework credited the child with less knowledge than was determined by the PCC metric. That generative phonology affords a child the maximum phonological knowledge again relates to the fact that this assessment examines both the production and function of sounds in the system. The PCC metric, as well as any other quantitative measure, only examines how sounds are produced at the phonetic level.

CONTINUUM OF PRODUCTIVE PHONOLOGICAL KNOWLEDGE		PERCENTAGE OF CONSONANTS CORRECT	
Type of Knowledge	Target Sounds	Severity Adjective	Target Sounds
Type 1 ("most") Adult-like representation No phonological rules	m n ŋ pb kg sz w j h	Mild PCC > 85%	m n ŋ k s w h
Type 2 Adult-like representation Optional phonological rules	td	Mild-Moderate PCC > 65%	p td g
Type 4 Adult-like representation in some, but not all positions	r	Moderate-Severe PCC > 50%	b
Type 6 ("least") Nonadult-like representation in all positions	fv θð ʃ tʃ dʒ l	Severe PCC < 50%	fv ð z ʃ tʃ dʒ j l r

Table 4. Continuum of productive phonological knowledge and percentage of consonants correct (PCC) for Child 6.

This observation has further implications for treatment. Returning to the case of Child 6, targets [b,z,j] would not be selected as potential treatment targets within a generative approach to pattern analysis because the child demonstrated type 1 knowledge of these sounds. [b,z,j] would likely be considered for treatment using the PCC metric since these sounds were characterized as moderately to severely involved. In some instances, sounds that do not require clinical attention may, in fact, be treated when PCC values are considered. Quantitative measures may overtarget and, therefore, overtreat sounds that a child already knows. There is preliminary experimental support (Gierut, 1985b) for granting a child the maximum phonological knowledge and for not directly treating those sounds that a child already knows; i.e., error sounds for which a child has relatively more knowledge have been shown to spontaneously improve without treatment.

Thus, the knowledge continuum and the PCC metric appear to be comparable in cases of sounds produced in error consistently and frequently (i.e., 0% accuracy/"least" knowledge). Here, the selection of treatment targets using the knowledge continuum and the PCC metric would be essentially equivalent. In all other cases, rankings on the knowledge continuum do not correspond to PCC values. PCC values seem to underestimate a child's productive phonological knowledge; consequently, sounds recommended for treatment based on PCC values would not likely be targeted for treatment based on the knowledge continuum.

Continuum of knowledge as an index of other selection factors

The continuum of knowledge may also correspond to other selection factors, e.g., ease of production, stimulability, typological markedness. With regard to ease of production, perhaps sounds that require greater motor control are ranked at the end of the continuum labelled "least" knowledge. To date, ease of sound production has not been firmly established (cf. Dinnsen, 1980; Locke, 1972; Ohala, 1980). In the absence of more definitive studies, it is difficult to identify those sounds that may require greater motor control.

With regard to stimulability, Dinnsen and Elbert (1984) suggested that sounds which are stimulable may be indicative of greater productive phonological knowledge than sounds which are not stimulable. This hypothesis has not yet been empirically evaluated.

The knowledge continuum may also reflect typological markedness. Typological markedness refers to a linguistic phenomenon which identifies a relationship among sounds, such that the occurrence of one sound in a language predicts or implies the occurrence of other sounds in that same language. The predicting or implying sound is "marked" relative to the predicted or implied sound, i.e., "unmarked." For example, if a language has voiced obstruents (e.g., stops, fricatives, and affricates), it will also have voiceless obstruents; voiced obstruents are marked relative to voiceless obstruents. Markedness relationships of this type have been identified by examining the sound systems of languages of the world (Greenberg, 1966; Greenberg, Ferguson, & Moravscik, 1978). Perhaps, sounds that are typologically unmarked are associated with "most" knowledge and sounds that are typologically marked are associated with "least" knowledge. This hypothesis remains open to empirical test.

Predictions about Learning

The final component in the assessment process involves making predictions about a child's learning. Predictions about learning provide a priori information about the extent to which a child's error pattern will be interrupted and restructured. Predictions about learning are primarily motivated by the generalization literature (see Elbert & Gierut, 1986, for review and discussion). For example, if a child produces a pattern of error involving the liquid [r] and is taught to produce [ɝ], it is likely that production of [r,ɚ] will also improve (Elbert & McReynolds, 1975; Hoffman, 1983). Predictably, this child's pattern of correct production would be reorganized to include all related allophones, [r,ɝ,ɚ].

There are two specific predictions that have consistently been observed across the different approaches to pattern analysis, including generative phonology. The first prediction is that untreated sounds within a pattern of error will be produced with some degree of accuracy following treatment of a particular sound within that pattern. In other words, generalization learning extends to untreated sounds within an error pattern. This type of learning has been observed, for example, within the place-voice-manner framework; teaching the manner of frication through production of [s] resulted in learning about frication in [z] (Elbert et al., 1967). Similarly, in the distinctive feature approach, teaching the [+strident] feature in the sound [f] resulted in accurate production of other sounds involving the [+strident] feature, [v,s,z,tʃ] (McReynolds & Bennett, 1972). Using a phonological process framework, elimination of the process of deletion of final consonants by teaching production of [p,d,s,θ] in final position resulted in the acquisition of other final consonants, [b,t,k,g,f,v,z,ʃ], affected by this same process (Weiner, 1981). Finally, teaching one sound ranked at a particular level of knowledge on the continuum resulted in accurate production of other sounds also ranked at this same level of knowledge (Gierut, 1985b). Across each of the different frameworks, then, generalization to untreated sounds within a specific pattern is a predictable type of reorganization that can be expected.

The second prediction is that production of treated sounds will be more accurate than production of untreated sounds within a pattern. Production of untreated sounds usually parallels, but lags behind, production of treated sounds. Generalization learning for treated sounds is, therefore, superior to generalization learning for untreated sounds. Again, this type of generalization has been observed across all pattern analyses (Costello & Onstine, 1976; Dinnsen & Elbert, 1984; Elbert et al., 1967; Gierut, 1985b; Hoffman, 1983; McReynolds & Bennett, 1972; McReynolds & Elbert, 1981b; Weiner, 1981). Reorganization of this type is, likewise, predictable.

Both of these predictions about learning provide information about how specific error patterns will restructure; they do not, however, provide an estimate of how a child's overall sound system will be affected. The effect that treating one pattern of production will have on a child's entire sound system is not predictable from most assessment frameworks. Generative phonology is the only assessment framework that provides empirical support for predictions about system-wide changes following treatment. Specifically, it has been demonstrated (Gierut, 1985b; Gierut et al., in review) that the starting point of treatment on the knowledge continuum is a predictor of the extent of system-wide reorganization and change. When treatment began with targets sounds ranked at "least" knowledge, extensive generalization to, and reorganization of, the entire sound system was observed. Production of untreated sounds ranked at "least" knowledge improved, in addition to production of other untreated sounds ranked at all higher levels on the knowledge continuum. When treatment was initiated at the opposite end of the continuum, "most" knowledge changes in the sound system were also observed; however, these changes were limited to only those sounds ranked at treated levels of knowledge (i.e., generalization within a pattern). System-wide reorganization was generally not observed when treatment began at "most" knowledge.

Predictions of this type have important clinical ramifications. The nature and extent of treatment is facilitated if both the degree of learning within an error pattern, as well as the degree of restructuring across the sound system can be projected. No other approaches to pattern analysis offer similar predictions about systematic restructuring of the overall sound system.

Discussion

The generative framework is a relatively new procedure of pattern analysis that presents information comparable, in many instances, to other assessments about the nature and treatment of speech sound disorders. As in other frameworks, a child's pattern of production is identified, sounds for treatment are selected on the basis of this pattern, and learning is projected. The generative framework for assessment, however, extends and qualifies information provided by other pattern analyses. Generative phonology, like other pattern analyses, characterizes the production of sounds in a child's sound system; it also describes the function of sounds and identifies the relationship among sounds. In selection of treatment targets, generative phonology relies on a hierarchical, ordered relationship among types of productive phonological knowledge. This criterion factor is different than those selection considerations used in other pattern analyses. Finally, changes in specific patterns of error are predictable from each assessment framework; generative phonology, moreover, projects how the overall sound system will be restructured. Therefore, the generative approach not only meets, but extends the basic elements of pattern analysis. Generative phonology adds to the three component assessment process and furthers the current state of information about the nature and treatment of functional speech sound disorders in children. This is not to say, however, that other non-generative pattern analyses should be abandoned or replaced. These frameworks have provided considerable insight into the assessment of speech sound disorders and serve as an impetus for further research.

For example, while it has been demonstrated that generative phonology offers a unique and different approach to assessment, it has not yet been determined whether this approach also provides the most accurate assessment. Single-subject treatment studies provide a suitable testing-ground for empirically evaluating predictions that derive from the different pattern analyses. If the claims and predictions of generative analysis are borne out in treatment, but those of other pattern analyses are not, then the generative approach would be judged most accurate in assessing speech sound disorders. On the other hand, if the claims and predictions of other pattern analyses are confirmed in treatment, but those of generative phonology are not, then the other pattern analyses would represent more accurate assessments of speech sound disorders.

The role of generative phonology in treatment also needs to be further established. Treatment methods common to other pattern analyses, e.g., minimal pair contrast treatment, have been used within the generative framework (Gierut, 1985b, 1986; Gierut et al., in review). Perhaps, different models or strategies of treatment specific to the generative approach will need to be developed. These procedures may, for example, take the form of morphophonemic treatment (Dinnsen, personal communication) where production of a sound is contrasted in phonologically relevant contexts using morphemically-related items, e.g., "pig" versus "piggy," "dog" versus "doggie." This approach to treatment may be particularly well-suited for a child displaying errors involving phonological rules operating on adult-like lexical representations (e.g., knowledge type 2). Treatment may also focus on category formation (Gierut, 1985b; Leonard & Brown, 1984) where specific sounds or lexical items are associated with new and narrower sound categories. This treatment strategy may be most appropriate for a child exhibiting errors that result from inventory constraints (e.g., knowledge type 6).

Finally, generative phonology may offer new insights into descriptions of normally developing sound systems. The claims of generative phonology, particularly those related to productive phonological knowledge, have not been examined with respect to young children's sound systems. It is not clear whether generative phonology is appropriate for use with these children given that the analysis procedures rely heavily on a child's knowledge of morphemes and the analyzability of words into morphemes. Whether young children segment, represent, process, or produce speech as distinct morpheme-sized units is a question open to considerable debate (cf. Peters, 1983).

In conclusion, generative phonology has direct applications in the clinical assessment process and can be differentiated from other approaches to pattern analysis on the basis of empirical claims. This framework also offers new directions for clinical research and remediation of disordered sound systems.

Endnotes

1 Several years ago, Compton (1970, 1975, 1976) reported using the generative framework in descriptions and treatment of children's speech sound errors. These descriptions involved radical modifications of generative phonology theory (Chomsky & Halle, 1968) that have not generally been accepted within theoretical linguistics. Moreover, the descriptions of sound errors developed by Compton cannot be distinguished from those developed within a phonological process approach, since both basically assume that a child maintains adult-like underlying representations of morphemes (see Maxwell, 1979, 1984, for a further discussion of this point).

2 Generative phonological descriptions of a child's sound system rely solely on production data. It has been suggested (Barton, 1978) that speech perception or discrimination data may provide information about a child's phonological knowledge. There are, however, some inherent difficulties in using data from speech perception or discrimination to evaluate phonological knowledge. For example, Locke (1980a, 1980b) has reported that it is difficult to accurately and adequately assess a child's perceptual skills. Also, the role of perception or discrimination in learning sounds during treatment has not been clearly established (Williams & McReynolds, 1975; Winitz, 1975). Furthermore, recent evidence from primary languages, normal language development, speech disorders, and second language learning suggests that speech production and speech perception may be independent processes (Dinnsen, 1985; Straight, 1980).

3 Although Elbert et al. (1984) used many terms and concepts particular to generative phonology, no generative analyses were developed or are otherwise available for these data. In fact, there has been no demonstration, to date, that it is appropriate to equate baseline scores with productive phonological knowledge.

4 A child's performance on baseline measures may not adequately or accurately establish productive phonological knowledge or a continuum of knowledge. However, a child's performance, i.e., generalization, during treatment does reflect his or her phonological knowledge (Dinnsen & Elbert, 1984; Elbert et al., 1984; Gierut, 1985b). That is, performance on sounds for which a child has "most" productive phonological knowledge is generally superior to performance on sounds for which a child has "least" knowledge. It appears that performance (baseline) may not be indicative of phonological knowledge, but that phonological knowledge is indicative of performance (generalization). The association between productive phonological knowledge and performance seems to be unidirectional.

References

Barton, D. (1978). *The role of perception in the acquisition of phonology*. Bloomington, IN: Indiana University Linguistics Club.

Bernthal, J.E., & Bankson, N.W. (1981). *Articulation disorders*. Englewood Cliffs, NJ: Prentice-Hall.

Camarata, S., & Gandour, J. (1984). On describing idiosyncratic phonologic systems. *Journal of Speech and Hearing Disorders*, *49*, 262-266.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Compton, A.J. (1970). Generative studies of children's phonological disorders. *Journal of Speech and Hearing Disorders*, *35*, 315-340.

Compton, A.J. (1975). Generative studies of children's phonological disorders: A strategy of therapy. In S. Singh (Ed.), *Measurement procedures in speech, hearing, and language* (pp. 35-92). Baltimore: University Park Press.

Compton, A.J. (1976). Generative studies of children's phonological disorders: Clinical Ramifications. In D.M. Morehead & A.E. Morehead (Eds.), *Normal and deficient child language* (pp. 61-96). Baltimore: University Park Press.

Costello, J., & Bosler, C. (1976). Generalization and articulation instruction. *Journal of Speech and Hearing Disorders*, *41*, 359-373.

Costello, J., & Onstine, J. (1976). The modification of multiple articulation errors based on distinctive feature theory. *Journal of Speech and Hearing Disorders*, *41*, 199-215.

Dinnsen, D.A. (1980). Phonological rules and phonetic explanation. *Journal of Linguistics*, *16*, 171-338.

Dinnsen, D. A. (1984). Methods and empirical issues in analyzing functional misarticulation. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph No. 22, pp. 5-17). Rockville, MD: ASHA.

Dinnsen, D.A. (1985). A re-examination of phonological neutralization. *Journal of Linguistics*, *21*, 265-279.

Dinnsen, D.A., & Elbert, M. (1984). On the relationship between phonology and learning. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph No. 22, pp. 59-68). Rockville, MD.: ASHA.

Edwards, M.L., & Bernhardt, B. (1973). Phonological analyses of the speech of four children with language disorders. Unpublished paper, Stanford University, Palo Alto, CA.

Elbert, M. (1985). From articulation to phonology: A change in perspective. *National Student Speech, Language, Hearing Association Journal*, *13*, 36-49.

Elbert, M., Dinnsen, D.A., & Powell, T. W. (1984). On the prediction of phonologic generalization learning patterns. *Journal of Speech and Hearing Disorders*, *49*, 309-17.

Elbert, M., Dinnsen, D.A., & Weismer, G., (Eds.) (1984). *Phonological theory and the misarticulating child* (ASHA Monograph No. 22). Rockville, MD: ASHA.

Elbert, M., & Gierut, J.A. (1986). *Handbook of clinical phonology: Approaches to assessment and treatment*. San Diego: College-Hill Press.

Elbert, M., & McReynolds, L.V. (1975). Transfer of /r/ across contexts. *Journal of Speech and Hearing Disorders*, *40*, 380-7.

Elbert, M., & McReynolds, L.V. (1978). An experimental analysis of misarticulating children's generalization. *Journal of Speech and Hearing Research*, *21*, 136-50.

Elbert, M., Shelton, R.L., & Arndt, W.B. (1967). A task for evaluation of articulation change: I. Development of methodology. *Journal of Speech and Hearing Research*, *10*, 281-8.

Fey, M.E., & Stalker, C.H. (in press). A clinical-experimental approach to treatment of a child with an idiosyncratic (morpho)phonological system. *Journal of Speech and Hearing Disorders*.

Foss, D.J., & Hakes, D.T. (1978). *Psycholinguistics: An introduction to the psychology of language*. Englewood Cliffs, NJ: Prentice-Hall.

Gierut, J.A. (1985a). Generative phonology: Clinical applications in speech pathology. *Innovations in Linguistics Education*, *3*, 152-167.

Gierut, J. A. (1985b). *On the relationship between phonological knowledge and generalization learning in misarticulating children*. Doctoral dissertation, Indiana University, Bloomington, IN. (Also distributed by the Indiana University Linguistics Club, Bloomington.)

Gierut, J.A. (1986). Sound change: A phonemic split in a misarticulating child. *Applied Psycholinguistics*, *7*, 57-68.

Gierut, J.A. (in press). On the assessment of productive phonological knowledge. *National Student Speech, Language, Hearing Association Journal*.

Gierut, J.A., Elbert, M., & Dinnsen, D.A. (1985). On characterizing phonological knowledge in disordered sound systems. *Research on speech perception* (Progress Report 11, pp. 205-234). Bloomington, IN: Speech Research Laboratory, Department of Psychology.

Gierut, J.A., Elbert, M., & Dinnsen, D.A. (in review). A functional analysis of phonological knowledge and generalization learning in misarticulating children. Manuscript submitted for publication to *Journal of Speech and Hearing Research*.

Greenberg, J.H. (1966). Synchronic and diachronic universals in phonology. *Language*, *42*, 508-518.

Greenberg, J.H., Ferguson, C.A., & Moravscik, E.A. (Eds.) (1978). *Universals of human language: Phonology*. Stanford, CA: Stanford University Press.

Hodson, B.W. (1980). *The assessment of phonological processes*. Danville, IL: Interstate Press.

Hoffman, P.R. (1983). Interallophonic generalization of /r/ training. *Journal of Speech and Hearing Disorders*, *48*, 215-221.

Ingram, D. (1981). *Procedures for the phonological analysis of children's language*. Baltimore: University Park Press.

Leonard, L.B., & Brown, B.L. (1984). Nature and boundaries of phonologic categories: A case study of an unusual phonologic pattern in a language-impaired child. *Journal of Speech and Hearing Disorders*, *49*, 419-428.

Locke, J.L. (1972). Ease of articulation. *Journal of Speech and Hearing Research*, *15*, 194-200.

Locke, J.L. (1980a). The interference of speech perception on the phonologically disordered child. Part I: A rationale, some criteria, the conventional tests. *Journal of Speech and Hearing Disorders*, *45*, 431-444.

Locke, J.L. (1980b). The interference of speech perception on the phonologically disordered child. Part II: Some clinically novel procedures, their use, some findings. *Journal of Speech and Hearing Disorders*, *45*, 444-468.

Locke, J.L. (1983). Clinical phonology: The explanation and treatment of speech sound disorders. *Journal of Speech and Hearing Disorders*, *48*, 339-341.

Maxwell, E.M. (1979). Competing analyses of a deviant phonology. *Glossa*, *13*, 181-214.

Maxwell, E.M. (1984). On determining underlying phonological representation of children: A critique of the current theories. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph No. 22, pp. 18-29). Rockville, MD: ASHA.

Maxwell, E.M., & Rockman, B.K. (1984). Procedures for linguistic analysis of misarticulated speech. In M. Elbert, D.A. Dinnsen, & G. Weismer (Eds.), *Phonological theory and the misarticulating child* (ASHA Monograph No. 22, pp. 69-84). Rockville, MD: ASHA.

McDonald, E.T. (1964). *A deep test of articulation*. Pittsburgh: Stanwix House.

McLean, J. (1970). Extending stimulus control of phoneme articulation by operant techniques. In F.L. Girardeau & J.E. Spradlin (Eds.), A functional approach to speech and language (ASHA Monograph No. 14, pp. 24-47). Rockville, MD: ASHA.

McReynolds, L.V. (1972). Articulation generalization during articulation training. Language and Speech, 15, 149-55.

McReynolds, L.V., & Bennett, S. (1972). Distinctive feature generalization in articulation training. Journal of Speech and Hearing Disorders, 37, 462-70.

McReynolds, L.V., & Elbert, M. (1981a). Criteria for phonological process analysis. Journal of Speech and Hearing Disorders, 46, 197-204.

McReynolds, L.V., & Elbert, M. (1981b). Generalization of correct articulation in clusters. Applied Psycholinguistics, 2, 119-132.

McReynolds, L.V., & Engmann, D. (1975). Distinctive feature analysis of misarticulations. Baltimore: University Park Press.

Ohala, J.J. (1980). The application of phonological universals in speech pathology. In N. Lass (Ed.), Speech and language: Advances in basic research and practice (Vol. 3, pp. 75-97). New York: Academic Press.

Olswang, L.B., & Bain, B.A. (1985). The natural occurrence of generalization during articulation treatment. Journal of Communication Disorders, 18, 109-129.

Peters, A.M. (1983). The units of language acquisition. New York: Cambridge University Press.

Prather, E.M., Hedrick, D.L., & Kern, C.A. (1975). Articulation development in children aged two to four years. Journal of Speech and Hearing Disorders, 20, 179-191.

Shelton, R., & McReynolds, L.V. (1979). Functional articulation disorders: Preliminaries to treatment. In N. Lass (Ed.), Speech and language: Advances in basic research and practice (Vol. 2, pp. 1-111). New York: Academic Press.

Shriberg, L.D., & Kwiatkowski, J. (1980). Natural process analysis: A procedure for phonological analysis of continuous speech samples. New York: Wiley.

Shriberg, L.D., & Kwiatkowski, J. (1982a). Phonological disorders I: A diagnostic classification system. Journal of Speech and Hearing Disorders, 47, 226-241.

Shriberg, L.D., & Kwiatkowski, J. (1982b). Phonological disorders III: A procedure for assessing severity of involvement. Journal of Speech and Hearing Disorders, 47, 256-270.

Singh, S. (1976). Distinctive features: Theory and validation. Baltimore: University Park Press.

Straight, H. S. (1980). Auditory versus articulatory phonological processes and their development in children. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Perception* (Vol. 2, pp. 43-71). New York: Academic Press.

Turton, C.J. (1973). Diagnostic implications of articulation testing. In W.D. Wolfe & D.J. Goulding (Eds.), *Articulation and learning* (pp. 195-232). Springfield, IL: Charles C. Thomas.

Weber, J.L. (1970). Patterning of deviant articulation behavior. *Journal of Speech and Hearing Disorders*, *35*, 135-141.

Weiner, F.F. (1979). *Phonological process analysis*. Baltimore: University Park Press.

Weiner, F.F. (1981). Treatment of phonological disability using the method of meaningful minimal contrast: Two case studies. *Journal of Speech and Hearing Disorders*, *46*, 97-103.

Williams, G., & McReynolds, L. V. (1975). The relationship between discrimination and articulation training in children with misarticulations. *Journal of Speech and Hearing Research*, *18*, 401-412.

Winitz, H. (1975). *From syllable to conversation*. Baltimore: University Park Press.

Wright, V., Shelton, R., & Arndt, W. (1969). A task for evaluation of articulation change: III. Imitative task scores compared with scores for more spontaneous tasks. *Journal of Speech and Hearing Research*, *12*, 875-884.

Appendix A

Three-minute portion of connected speech used to
calculate PCC values for Child 6

ɪts ə tisɚ vaɪt oʊə dɛr

it's a teacher right over there

dɛm ə gʊɪʃ

them the girls

dɛm dɪts̪ ʌp ən daʊn

them gets up and down

dɛm siz dɛm baɪtɪn

them sees them writing

dɛm aʊt baʊt vaʊnz ən kʌɾɚ

them about about crayons and color

vaʊns ən vaɪɾɪn

crayons and writing

vaʊnts̪

crayons

ɪts ə kaɪnə maɪk ɪt baɪk ə haʊs

it's a kinda make it like a house

dɛm tɛl hɪm sɛt daʊn

them tell him sit down

tɛl əm haʊ ən dɛn haɪm ə kʌɾɚ

tell him how and then time to color

201

dɛm də ʌɾɚ gaɪ ɪn ə dark

them the other guys in the dark

hɪm poʊsə kʌm ɪnsaɪd

him supposed to come inside

əkʌz hɪm wɛr weɪn waɪk sækət͜s ən sæməs

because him wear rain like jackets and pajamas

aɪ doʊnt noʊː

I don't know

wɛoʊ wi du dæt ɛvɚi taɪm

well we do that every time

wi ʤʌsə dʉ dæt bʌ wi gɛə hɛr ʌp ən du dæt

we just do that but we dress her up and do that

hɪm seɪ wɛn ɪt gɛt darkɚ

him say when it gets darker

dɛn hɪm goʊ aʊtsaɪd ɪn goʊ stɔr

them him go outside and go store

mɛn ər kʊkɪn ən meɪkɪn sʌm dæt ən də æɾɚ gɚz ən sʌmə

men are cooking and making some that and the other girls and some of

swɪks sʌm sʌpɚ

fixes some supper

toʊst ən weɪbi ən sʌm mʌswums

toast and gravy and some mushrooms

sʌm əv wæt æn sʌm sɪkən

some of bread and some chicken

ə kʌz dɛr meɪkɪn əbʌsəs sʌpɚ

because they're making everybody's supper

æn dɔ bʌ də ʌʃɚ gɛr se̯ɪɪn doʊnt meɪk ɛni dæp dæs

and the but the other girls saying don't make any (unintelligible)

Effects of Talker Uncertainty on Auditory Word Recognition: A First Report*

John W. Mullennix and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*The research reported in this paper was supported, in part, by NIH Research Grant NS-12179, and, in part, by NIH Training Grant NS-07134 to Indiana University in Bloomington. We would like to thank Paul A. Luce for the helpful discussions, comments, advice, and programming assistance he provided throughout the entire study.

Abstract

Individual talkers vary in source and vocal-tract configurations, loudness, speaking rate, and accent among other variables. The effects of these differences have largely been ignored in research concerning auditory word recognition. In the present study, we investigated the effects of talker uncertainty on auditory word recognition using an identification task. Subjects were presented with monosyllabic CVC English words presented at various signal-to-noise ratios, and were required to identify them on an absolute basis. The single-talker group received words from one talker only, while the mixed-talker group received items from 15 different talkers. In addition, the effects of lexical structure upon identification were examined. Lexical density was manipulated within subjects by creating high and low density stimulus items as a function of number of neighbors for each word. The results showed that identification performance was worse under the mixed-talker conditions than the single-talker condition. Performance under both lexical density conditions was not significantly different, although an interaction was present between S/N ratio and density. The results suggest that variability due to talker uncertainty may be an important variable that affects the early stages of auditory word recognition. The absence of a main effect of lexical density suggests that perceptual processes related to talker normalization may logically proceed access to the lexical representation of words in long-term memory. The interaction between S/N ratio and density suggests that lexical structure is an important variable affecting auditory word recognition.

Effects of Talker Uncertainty on Auditory Word Recognition:

A First Report

The processes involved in recognizing spoken words are extremely complex. Information at sensory, phonetic, lexical, and higher-order contextual levels is used to arrive at the final word percept. Although alternatives to this view exist (i.e. Klatt, 1980), much work has suggested that analysis at multiple levels of processing contributes to the recognition of spoken words. The components which are intrinsic to the word recognition process may be organized in a fashion which is interactive (Elman and McClelland, 1984), hierarchial (Forster, 1976; 1979), or some combination thereof (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980). Regardless of how the processing components are organized, the manipulation of certain variables within an experiment can elucidate the contribution of hypothetical processing factors relevant to the word recognition process.

The experiment to be described here concerns an important area that has been neglected as a factor in auditory word recognition research, namely, the effects of talker differences on recognition. Individual talkers vary in a multitude of differing articulatory and acoustic characteristics, i.e., source and vocal tract configuration, vocal amplitude, speaking rate, accent, dialect, etc. For the average listener, these differences between talkers appear to be of little consequence in recognizing spoken words and understanding spoken language. This form of normalization or compensation has been viewed as a form of perceptual constancy, where differences in articulatory and acoustic composition of particular spoken items do not appear to overtly affect perception. Although subjectively it may appear as though differences attributed to different talkers do not have perceptual consequences, this is not necessarily the case. It is possible that the variability manifested by differences in articulatory and acoustic parameters attributed to talkers may be compensated for by a processing mechanism which quickly and efficiently "normalizes" the differences between talkers. In particular, the processes which are inherent components in recognition of spoken words may include such a mechanism. In the present study, the role of talker uncertainty was examined in a task which measured performance in recognizing spoken words. Before proceeding, however, background research will be summarized which pertains to the effects of talker variation upon speech and word perception. These studies formed the motivation for the work we carried out on word recognition.

A number of studies have looked at the effects of vocal tract differences on the perception of speech. In particular, a number of studies have focused on vowel perception (Summerfield and Haggard, 1973; Summerfield, 1975; Nearey, 1977; Assman, Nearey, and Hogan, 1982; Verbrugge, Strange, Shankweiler, and Edman, 1976; Strange, Verbrugge, Shankweiler, and Edman, 1976). These studies have used various tasks to examine performance for vowel perception under conditions where the vowels originated either from one talker or from a variety of talkers. Summerfield and Haggard (1973; Summerfield, 1975) used sets of synthetic stimuli in which spectral information was varied by manipulating values of F0 and vowel formant frequencies (i.e. creating the items "bed", "bird", and "bored"). Essentially, a number of stimuli were created which varied in formant components which emulated different vocal tracts. Using a choice RT task for the vowels, Summerfield and Haggard (1973) discovered that RTs were slower for a particular target syllable when the target was preceded by a syllable which contained spectral parameters specifying a different vocal tract than the target stimulus. They concluded that the slower RTs to the target in this situation reflected the time taken for the operation of a vocal tract normalization process, although the details of this mechanism were left unspecified.

A slightly different approach to studying the effects of talker variability on vowel perception was taken by Strange and her colleagues (Strange et al., 1976; Verbrugge et al., 1976). In a series of experiments, the effects of talker variation and consonantal context on vowel identification were examined. In one experiment, Strange et al. compared performance in terms of identification responses for naturally-produced isolated vowels and for medial vowels produced in a /p-p/ context. Two conditions were present: a segregated-talker condition, in which subjects were given stimuli from a single talker, and a mixed-talker condition, in which subjects were given stimuli randomly selected from 15 different talkers (males, females, and children). Strange et al. found that for both isolated vowels and /p-p/ vowels, identification performance was worse in the mixed-talker case. Thus, uncertainty due to talker variability had a detrimental effect on vowel identification.

In another experiment, Strange et al. (1976) again compared segregated versus mixed-talker conditions, with the vowels produced in a /C-C/ context, where the consonants randomly varied between six stop-consonants /b,d,g,p,t,k/. Here, they found no difference in performance between groups. Strange et al. compared the results to the isolated vowels and /p-p/ context vowels in terms of the relative usefulness of consonantal context cues in vowel identification. However, they did not offer an explanation for the lack of an effect of talker variability on identification when the context was in a /C-C/ form.

In considering the Strange et al. results, a couple of caveats must be kept in mind. First, as Assman, Nearey, and Hogan (1982) pointed out, the results in these particular experiments could have been confounded by dialect discrepencies between talkers and listeners, as well as orthographic interference within the response procedure. Assman et al. (1982) examined vowel identification for segregated and mixed-talker conditions, where these two factors were controlled. They used natural, isolated vowels and gated vowels (natural vowels with only the center portion remaining). For both types of vowels, performance was worse in the mixed-talker case than in the single-talker condition. Thus, it appears that at least for isolated and gated natural vowels, variability due to talkers has a detrimental effect upon vowel identification.

In addition to the work done on the effects of talker variability on vowel perception, attempts have been made using somewhat different experimental paradigms to examine talker effects at the word level. Cole, Coltheart, and Allard (1974; Allard and Henderson, 1975) used an auditory analog of the Posner and Mitchell (1967) same-different RT task, which investigated physical and name codes. Cole et al. (1974) used a restricted set of CV alphabetic-name stimuli and isolated vowels. The stimuli were produced by one male and one female speaker. For "same" responses, two possibilities existed: First, a physical identity match on the basis of same word and voice could result in a same response, i.e. a male "bee" and a male "bee". Secondly, a name match in which voice was irrelevant could also result in a "same" response, i.e. a male "bee" and a female "bee". Cole et al. found that RTs for "same" responses where the stimuli differed in voice were slower than those found for same-voice "same" responses. However, a similar pattern for stimuli in the same and different voices also existed for "different" responses. That is, RTs for the "different" responses were faster with pairs of stimuli occurring in the same voice. As this pattern of RTs was different than what Posner and Mitchell (1967; Posner, Boies, Eichelman, and Taylor; 1969) had found in letter matching, the results were interpreted as evidence against a direct auditory analog of visual physical and name codes.

In another study, Allard and Henderson (1975) found identical results for a set of CVC words. However, two additional aspects of the Allard and Henderson (1975) experiment were interesting: First, the effect of voice on same RTs disappeared over sessions as a result of practice; and, second, the voice effect was only manifested in the subjects who were slower responders. It is unclear in this case as to whether experience with the voices over time resulted in a "readjustment" of the perceptual mechanisms for responding to the stimuli, or whether it was simply due to a change in task strategies. It is possible that faster responders may have been more adept at developing a response strategy, hence showing no effect of voice for "same" decisions. Performance differences may have been due to task strategies which were irrelevant to processes affected by talker variation. In addition, the stimuli used in both studies were highly constrained: Cole et al. (1974) used 4 CVs and 4 vowels, while Allard and Henderson (1975) used only 5 words. It is possible that with such a small number of stimuli, subjects were capitalizing upon aspects of the stimuli which, under normal conditions, would not be attended to. Thus, these results must be viewed with caution when interpreting them in terms of their effects upon perception. Nevertheless, there is some suggestion from these studies that talker uncertainty may affect auditory word perception.

The effect of speaker's voice at the word level has also been investigated within more traditional recognition and recall paradigms within the human memory literature. Craik and Kirsner (1974) conducted a series of experiments which examined recognition performance in continuous-string auditory word lists. Subjects were required to listen to a long, continuous list of common nouns. The words were spoken either by a male or female talker. The subjects performed a recognition task, i.e., for each word presented they judged it as "old" or "new". The critical variable was whether the word, when repeated the second time, was in the same or different voice as the original word. The results showed that recognition performance was better when the word was repeated in the same voice rather than a different voice. However, the ability to recall the voice in which the original word occurred was not affected by the voice in which the word was repeated. Thus, recognition memory performance for words was affected by variation in voice of the word, but recall was unaffected, at least within the traditional experimental paradigm used to study recognition memory and free recall.

One recent experiment by Mattingly, Studdert-Kennedy, and Magen (1983) explicity tested the effects of speaker and dialect variation on memory. In a serial-ordered recall experiment using digits, Mattingly et al. constructed three conditions for recall: (1) a "single-speaker condition", in which all the words came from one speaker only; (2) a "mixed-speaker, same dialect" condition, in which the words came from three English speakers with the same dialect, and (3) a "mixed-speaker, different dialect" condition, in which the words came from three different speakers, each speaker having a different dialect. Their results indicated that only performance in the primacy region of the serial position curve was worse for the mixed different-dialect condition than for the single-speaker and mixed same-dialect conditions, which did not differ from one another. Mattingly et al. (1983) interpreted these results as suggesting that phonological uncertainty affected either rehearsal or encoding processes in memory, such that items spoken in varying dialects were harder to remember. Under these conditions, the variation in speaker characteristics did not cause any observable performance deficits.

The pattern of results found in the Mattingly et al. experiment is remarkably similar to work done in our laboratory on the recall of lists of synthetic speech (see Luce, Feustel, and Pisoni, 1983). Luce et al. (1983) found that serial-ordered recall of lists of natural and synthetic words resulted in performance decrements in the primacy region of the serial position curve for synthetic speech. This finding was interpreted as evidence for the claim that synthetic speech incurs greater demands in processing, such that rehearsal and/or encoding processes suffered. The results of Mattingly et al. (1983) are also related to findings reported by Rabbitt (1968), who found that serial recall of digits in the first half of a list of words suffered more when the latter half of the list was presented in noise, regardless of whether the first half of the list was presented in noise or not. The findings of Rabbitt (1968) have been explained in terms of increased processing demands due to presentation of items in noise interfering with rehearsal/encoding processes in memory for earlier items in the list. The results are analogous to processing demands of synthetic speech affecting recall (Luce et al., 1983), and processing demands of phonological uncertainty affecting recall (Mattingly et al., 1983).

Taken together, the experiments that have been conducted examining effects of talker variability on vowel perception and word recognition and recall suggests several conclusions. On the one hand, the results found by Summerfield and Haggard (1973), Summerfield (1975), and Assman et al. (1982) suggest that at the segmental acoustic-phonetic level, talker variability incurs a decrement in perception. Strange et al. (1976) found similar results for vowels in isolation and /p-p/ contexts, but found no effect with vowels in a /C-C/ context. Some evidence is provided for talker effects in the Posner-type tasks (Cole et al., 1974; Allard and Henderson, 1976), but as mentioned above, these findings are obscured by possible response strategies using that particular task. Voice differences do have effects on recognition of words in a continuous string (Craik and Kirsner, 1974), but do not affect serial-recall unless the dialect is also different (Mattingly et al., 1983). At the very least, this body of research which has studied talker uncertainty as an experimental variable suggests the possibility that the uncertainty due to a change in talker may have both perceptual and processing consequences. Unfortunately, the nature of these differences is not well understood nor are there sufficient data in the literature to permit one to formulate a well reasoned account of these differences.

With regard to word recognition processes, an appropriate task is needed which will focus directly on these questions. The aim of the present study was to study the effects of talker uncertainty on the identification of isolated spoken words. By examining performance differences in identification under conditions of exposure to words from one talker or from many talkers, inferences can be made about underlying processing operations involved in the word recognition process. If talker variability has a detrimental effect on identifying spoken words, it is possible that an underlying mechanism may be used for transforming different-talker input into a more abstract form as it can be used by the perceptual system.

The studies discussed above all used extremely restricted sets of synthetic and natural stimuli. Evidence exists that synthetic speech requires greater capacity and incurs greater encoding demands than natural speech (Luce et al., 1983; Pisoni, Nusbaum, and Greene, 1985). Thus, the use of synthetic speech in exploring spoken word recognition may not give an accurate picture of what happens when natural speech is heard. Also, with a restricted set of stimuli within a given task, responses may be based upon aspects of the stimuli which under normal conditions may not be attended to by listeners.

In the present experiment, natural speech was used with a reasonably large set of words. The responses in the identification task were unconstrained by the use of an open response set. Subjects were required to identify the word which was presented on each trial. Hence, the experimental situation more closely resembles conditions which may be found in a more "naturalistic" situation, where naturally-produced words are heard and identified in the course of normal conversation. However, the lack of sentential context precluded the use of higher-order contextual information (i.e. syntactic, semantic, pragmatic) in making responses. Thus, any observed effects may be relegated to processes related to word recognition and lexical access or earlier analyses of the acoustic-phonetic information in the signal.

In addition to talker uncertainty, another factor was manipulated in this study. This factor is related to the structure of words in the mental lexicon. Landauer and Streeter (1973) and Eukel (1980) have suggested that lexical structure may affect processes of word recognition and lexical access. Recently, Luce (1985; 1986), using auditory and visual word recognition tasks, has found that several structural factors which include measures of lexical density and similarity are important in accounting for word recognition performance. In the present experiment, the density of the lexical space of words was manipulated. One particular measure (the number of words differing from a given lexical item by one phoneme) provided an index for a given word in regards to its position in lexical space. Words of high lexical density and low lexical density were selected in order to study the effects of lexical similarity on word recognition. Thus, a situation was created where both the effects of lexical structure and talker variability could be studied.

The basic predictions for the present experiment were as follows: First, if uncertainty due to talker variability has a detrimental effect on auditory word recognition, then performance should be worse under conditions where subjects received stimuli from many talkers rather than just one talker. Second, if lexical density has an effect on auditory word recognition, performance should differ as a function of the density condition (high or low). An interaction or lack thereof between lexical density and talker uncertainty should allow statements to be made about the hypothetical locus of a talker-normalizing perceptual mechanism. If both variables have effects on performance and do not interact, this outcome would suggest that the manipulations affect processes at different processing levels. If both variables have effects and do interact, this would suggest that the processing operations affected by both manipulations may be occurring at some common locus.

Identification performance was studied at three different signal-to-noise ratios and over successive blocks of trials in the experiment. We predicted that overall performance would be better at higher S/N ratios, simply as a function of discriminability of the signal from the background noise. The pattern of performance over blocks provides information as to whether practice (or experience) has a differential effect as a function of talker condition. If experience with only one talker (versus many talkers) results in a relatively greater increase in performance over the blocks of trials, this would provide further evidence that a mechanism exists which "tunes in" on the talker characteristics to assist perceptual processing and facilitate word recognition processes.

Method

Subjects. Thirty-seven undergraduate students from an introductory psychology course at Indiana University were used as subjects. Fifteen subjects served as talkers to produce the stimuli, and 22 subjects served as listeners in the perceptual experiment. Each subject participated in one 1-hour session, and received partial course credit for participating in the experiment. All subjects were native speakers of English with no history of a speech or hearing disorder.

Stimuli. The stimuli consisted of 72 naturally spoken words obtained from each speaker. The words consisted of CVC monosyllabic English words that varied randomly in their consonants (i.e. stops, fricatives, affricates, liquids, and nasals) and vowels. The set of 72 stimuli was identical across all speakers. The stimuli were recorded on audiotape from each speaker in a sound-attenuated IAC booth using an Electro-Voice Model D054 microphone and a Crown 800 series tape recorder. Each stimulus to be recorded appeared on a CRT screen in front of the subject, embedded in the carrier sentence "Say the word ___ for me", where the blank corresponded to a particular target word. The talker was instructed to read the entire sentence aloud in a normal voice and at a normal speaking rate. Utterances were recorded from 7 male talkers and 8 female talkers. The sentences were converted to digital form via a 12-bit analog-to-digital converter at a 10,000 kHz sampling rate, and were low-pass filtered at 4.8 kHz. The words were then digitally edited from the carrier sentences to produce the final stimuli used in the study. Amplitude levels among words were equated using a software package designed specifically for digitally manipulating amplitude.

The test words were selected to differ in terms of lexical density using an on-line lexicon database consisting of a version of the 20,000 entry Webster's Pocket Dictionary. This database was used to compute a lexical density measure for each stimulus based on neighborhood similarity (see Luce, 1985; 1986). The measure of lexical density used here was defined as the number of neighbors (words differing by one phoneme from the stimulus) existing for a particular word. Low-density words had a value of 10 or less; high-density words had a value of 15 or greater. Thirty-six words were selected for each condition, resulting in a total of 72 stimuli.

The final constraint used in selecting words was related to familiarity. Familiarity ratings on a scale from 1 (unknown) to 7 (familiar and well-known) were obtained from data collected in a previous study by Nusbaum, Pisoni, and Davis (1984) for the lexical database used here. The stimuli selected for the present study met a 95% criterion of familiarity (translating to 6.65 on the rating scale). All 72 stimuli were rated at 6.65 or above. Thus, all the target words were rated as highly familiar to the subjects. This manipulation was done to insure that subjects were familiar with the words used in the experiment.

Procedure. Three experimental factors were manipulated: Talker variability, lexical density, and signal-to-noise (S/N) ratio. Talker variability was manipulated as a between-subjects factor, forming two groups with 11 subjects each. The single-talker group received stimuli from one talker only throughout the session, while the mixed-talker group received words from all 15 of the talkers. In the mixed-talker group, 5 words were randomly selected for presentation from 12 of the speakers, and 4 words from 3 of the speakers. In the single-talker group, each subject received the 72 stimuli from only one of the 15 different talkers. In other words, each subject in this condition received stimuli coming from a different speaker than each of the other subjects had received. Manipulation of the lexical density factor created two within-subject conditions: High-density and low-density target words. As previously mentioned, the low-density items consisted of words having fewer neighbors, while high-density items had more neighbors. Each subject received both types of items. Finally, the last variable manipulated was the signal-to-noise ratio. Each word was presented at 3 differing S/N ratios: +10 dB, 0 dB, and -10 dB. S/N ratio was manipulated within subjects. For all S/N conditions, the background of noise remained constant at 70 dB SPL, with the stimulus signal attenuated at 80 dB SPL, 70 dB SPL, and 60 dB SPL for the respective conditions.

The experimental paradigm involved the use of a word identification task. First, each stimulus was embedded in noise and was presented binaurally over matched and calibrated TDH-39 headphones to subjects. Subjects in the single-talker group were run individually, while subjects in the mixed-talker group were run in small groups varying from 1 to 4. For each trial, the subjects were instructed to identify the word that was presented in noise, and then type in a response on a CRT terminal. A prompt appeared on the CRT screen immediately after presentation of the stimulus to indicate that a response should be initiated. Subjects were instructed to type in a word corresponding to what they thought they had heard. They were not given any indication as to what words to expect during the experiment, except that they would be English words. After all the subjects had responded, a message appeared on the CRT indicating that the next stimulus would be presented. A 2-second ISI intervened between presentation of the message and the subsequent stimulus onset.

Three separate blocks of 72 trials were run, with approximately a 2-minute rest period between each block. Each word was presented only once in each block, with the word presented at a different S/N ratio in each particular block. For example, the word "batch" may have been presented at a -10 S/N ratio in the first block, at a 0 S/N in the second block, and at a +10 S/N in the third block. Within each block, words occurred randomly at all three of the S/N ratios, with one-third of the words presented at each S/N ratio in each block. The assignment of S/N ratio to each word was randomized, as was presentation of words within each block. Stimulus output and data collection were controlled on-line by a PDP-11/34a computer. Stimuli were output via a 12-bit digital-to-analog converter at a 10,000 kHz sampling rate, and were low-pass filtered at 4.8 kHz.

Results

The data were collected and analyzed on computer. Responses were analyzed in terms of percent correct identification. Table 1 displays the results for the mixed and single-talker conditions for high and low lexical-density, at the three S/N ratios, for each of the three blocks of trials. These data are also presented graphically in Figure 1.

Insert Table 1 and Figure 1 about here

An examination of the data shown in Figure 1 suggests the presence of several experimental effects. To quantify these observations, a four-way ANOVA was run with the factors of talkers (single or mixed), density (high or low), S/N ratio (+10, 0, or -10) and block (1st, 2nd, or 3rd block of trials).

Table 1

Identification performance of words
(percent correct)

SINGLE-TALKER GROUP

	Block 1	Block 2	Block 3
High Density			
+10 S/N	68.2	78.0	53.4
0 S/N	31.8	52.3	50.8
-10 S/N	3.0	3.8	12.9
Mean	34.3	44.7	39.0
Low Density			
+10 S/N	59.9	74.3	75.8
0 S/N	48.5	45.2	51.2
-10 S/N	3.8	10.5	7.3
Mean	37.4	43.3	44.8

MIXED-TALKER GROUP

	Block 1	Block 2	Block 3
High Density			
+10 S/N	65.9	81.1	39.4
0 S/N	15.2	43.9	46.8
-10 S/N	0.7	1.5	8.4
Mean	27.3	42.2	31.5
Low Density			
+10 S/N	56.4	36.4	75.0
0 S/N	38.6	52.7	29.3
-10 S/N	0.7	3.0	15.0
Mean	31.9	30.7	39.8

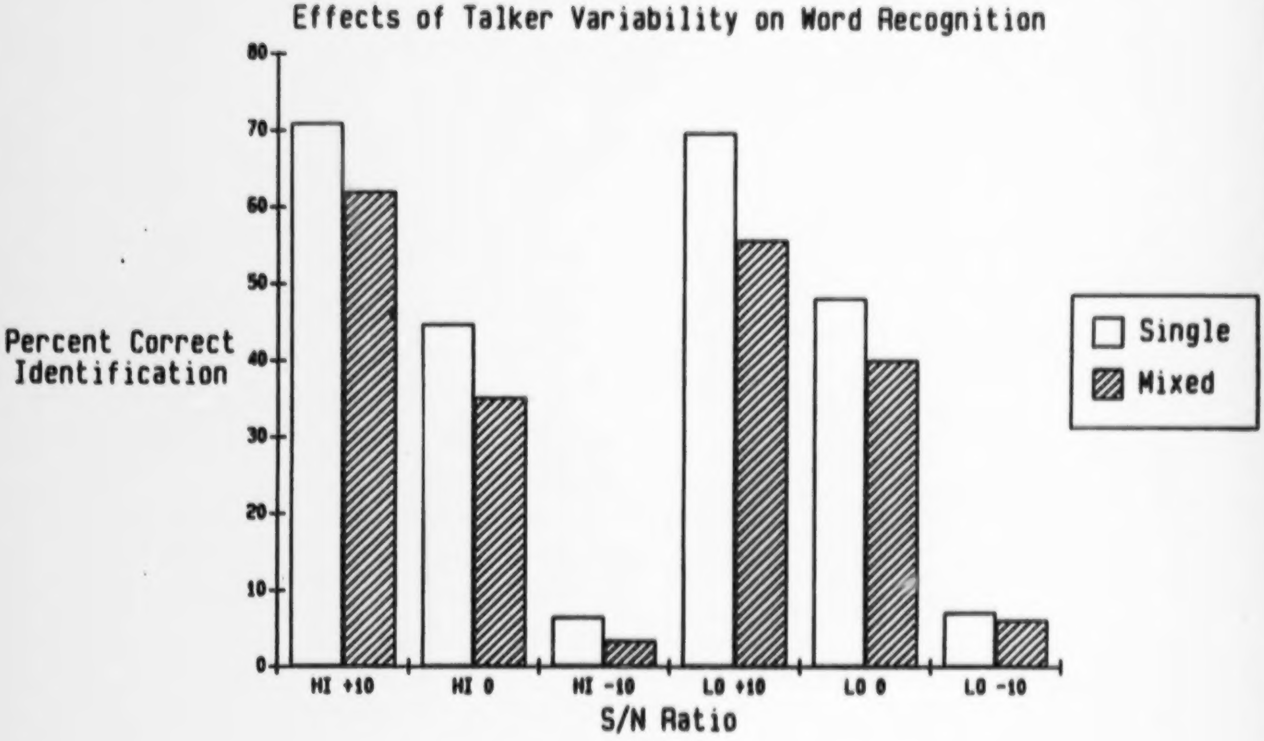


Figure 1. Percent correct identification of words as a function of talker group, density, and signal-to-noise ratio.

Three significant main effects were found. First, an effect of talker was present (F[1,20] = 8.26, $p < .01$). Identification was better for the single talker condition compared to the mixed talker condition (40.6% correct and 33.9%, respectively, averaged over all conditions). This result suggests that the uncertainty due to multiple talkers had a detrimental effect on auditory word recognition; words were not identified as well when stimuli came from many talkers.

Second, a main effect was found for S/N ratio (F[2,40] = 592.5, $p < 0.01$) Identification performance was best in the +10 S/N condition, next to best in the 0 S/N condition, and worst in the -10 S/N condition (see Figure 1). As expected, performance was a function of the discriminability of the speech signal from the background noise, with better identification at higher S/N ratio.

Finally, a main effect of block was observed (F[2,40] = 20.1, $p < 0.01$). Performance in the first block of trials was worse than the second and third block. Newman-Keuls tests showed that the second and third block did not differ reliably from one another, while the first block was significantly worse than the other two. This result suggests that experience with the stimuli obtained in the first block led to better performance in the latter blocks.

A number of significant interactions were also observed in these analyses. First, the two-way interactions will be considered. A density x S/N interaction was found (F[2,40] = 3.64, $p < .04$). Performance was slightly better for high-density items in the +10 S/N condition, and slightly better for low-density items in the 0 S/N and -10 S/N conditions. Newman-Keuls tests showed that these differences between density conditions at each S/N ratio were not significant. A S/N x talker interaction was found (F[2,40] = 4.03, $p < .03$). Newman-Keuls tests showed that performance was better for the single-talker group than the mixed-talker group in the +10 S/N and 0 S/N conditions, but not for the -10 S/N condition. A density x block (F[2,40] interaction was = 7.79, $p < .002$) present, with performance significantly different between the density conditions in the second block only (with performance higher for the high-density items). An interaction was also found for S/N x block (F[4,80] = 5.92, $p < .001$). Performance was significantly different within each block of trials for the three S/N conditions.

Three significant three-way interactions were also found. An interaction for density x block x talker was present (F[2,40] = 5.2, $p < .001$), S/N x block x talker (F[4,80] = 4.42, $p < .003$), and density x S/N x block (F[4,80] = 14.16, $p < 0.0$). The four-way interaction was also significant (F[4,80] = 7.76, $p < 0.0$).

Discussion

The results of the present experiment have implications for understanding the effects of both talker uncertainty and lexical density on auditory word recognition. First, the fact that performance was worse in the mixed-talker condition demonstrates that uncertainty due to talker variability has a detrimental effect on recognition. Words presented at +10 S/N and 0 S/N ratios were identified better when the words were from one talker, rather than 15 talkers. Performance for words presented at -10 S/N for both talker conditions is about the same, and probably reflects floor effects.

With regard to lexical density, no overall significant difference in performance between high and low density items was observed, although density entered into several interactions with the other variables. Under these particular conditions, it appears as though the density measure which was manipulated does not have any direct effect upon word recognition. However, the interaction between density and S/N ratio suggests that lexical structure in terms of the density manipulation may be important. Performance at the +10 dB S/N level was slightly better for high-density items. But, at the 0 dB S/N level, performance was better for the low-density items. At the -10 dB S/N level, performance was a little better for low-density items although there may have been a floor effect in performance at that level. These particular results suggest that when the acoustic-phonetic level of information becomes increasingly degraded by an increase in S/N level, the information from the lexical level of processing is relied on to a greater extent in order to identify the word item. Since low-density items have fewer neighbors to compete with (thus being less confusable), the low-density words are identified more often.

There are a number of possible reasons as to why performance did not directly vary as a function of density. First, it may be the case that the measure which we used was much too crude. Luce (1985) found that for visually presented words, lexical density measured in number of words differing from the target in one letter had a significant effect upon word identification. However, for auditorily presented words (Luce, 1986), the picture appears to be more complicated. With spoken words, acoustic-phonetic confusibility combines with a mulitiplicity of lexical structural components to contribute to spoken word recognition. Lexical density as measured by the absolute number of neighbors does not appear to be an overtly salient factor in recognition performance (see Luce, 1986). Hence, with a set of stimuli restricted to CVCs, this structural effect may not be displayed very robustly.

Second, there may not have been sufficient power to pick up any effects of the density measure. Each subject only made three responses to each word, one response at each S/N ratio. The magnitude of the effect of this particular density measure may be small, so that an insufficient number of observations per stimulus may preclude eliciting the effect. Although both of these explanations for the absence of an effect of density as used here are speculative, it may be useful to investigate lexical structure with a different metric in future studies of auditory word recognition.

With regard to the other findings, the effect of S/N ratio was anticipated. Performance was best at the highest S/N, and worst at the lowest S/N. The effect of block showed that performance was worst in the first block of trials, and was significantly better in the second and third blocks, with performance in the second and third blocks approximately the same. It appears as though practice under these conditions results in an improvement which reaches a ceiling by the end of the second block of trials. The lack of an interaction between the talker variable and blocks shows that performance over time was about the same for both talker groups.

Finally, the other various interactions which were found had little bearing upon the variables of interest (talker and density). No two-way interaction was found between talker condition and density; although a three-way interaction was found between talker, density, and block. Closer analysis showed that performance between talker groups differed as a function of density in the second block of trials.

The effects of talker uncertainty on auditory word recognition in the present experiment suggest that the processes which are involved in taking speech input and transforming it into a lexical representation must include mechanisms that adjust for differences between talkers. Whether this mechanism is a "vocal tract normalization" process (Summerfield and Haggard, 1973), or a higher-level mechanism which is responsive to uncertainty in a more generic sense is not clear at this time. The absence of effects due to lexical density precludes any speculation as to whether this putative mechanism occurs at a level which incorporates structural factors of the lexicon or not.

A number of questions arise as to the nature of this phenomenon, in particular, what aspects of word recognition processes may be affected by talker uncertainty. To further investigate questions of this sort, converging evidence from other experimental paradigms may be useful. For example, experiments which look at word recall from memory may be useful in determining whether talker variability affects processes relevant to memory. Mattingly et al. (1983) did not find any differences in performance between word lists that came from either one talker or many talkers; however, they used only three different talkers. It is very likely that three talkers will not induce enough variation in order to exhibit any effects upon recall. Also, the serial-recall procedure as they used it may not tax processing capacity to the extent that effects due to talker variability would be displayed robustly. It may be necessary to increase the processing load in the task (i.e., use a memory pre-load, increase presentation rate, etc.) before such effects would be exhibited. If talker variability in such a task has a detrimental effect upon recall, examination of the serial position curve could provide useful information concerning the locus of the effects. A lowered primacy region would indicate that either encoding or rehearsal processes are being affected by talker variation. Changes in the recency portion of the curve may not be as easily interpreted in terms of short-term memory (see Baddeley and Hitch, 1977; Bjork and Whitten, 1974; Greene, 1986a,b).

In considering the effects of lexical structure on word recognition, it may be more useful to use a confusability metric which accounts for a greater proportion of performance instead of the simple density measure as used in the present experiment (see Luce, 1986). If an interaction exists between a lexical density metric and a talker variable, this would suggest that one locus of the processes affected by talker variation is related to processes associated with accessing words in the lexicon. Conversely, if there is an effect of both lexical and talker variables with no interaction, then the processes affected by talker variation may exist at separate levels independently of structural factors in the lexicon.

In summary, the present investigation was designed to study the effect of talker uncertainty on auditory word recognition using English words that differed in lexical density. The result demonstrated consistent differences in identification performance of words in noise for single-talker, homogenous conditions over mixed-talker conditions. The uncertainty due to changes in talker from trial to trial in the mixed condition affected performance at the two most favorable S/N ratios. However, no effect of the lexical density manipulation was observed for the set of highly familiar CVC monosyllabic words used in the study. Obviously, further research will be necessary to understand the precise nature of these differences and to identify the locus of the perceptual mechanisms responsible for talker normalization effects in auditory word recognition. In future work it will be necessary to dissociate effects due to phonetic processing at early stages of language understanding from processes related to access of words in the mental lexicon. From the present result, it appears that the effects of talker normalization are primarily perceptual in nature and restricted to developing some segmental representation of the input signal.

References

Allard, F., and Henderson, L. (1976). Physical and name codes in auditory memory: The pursuit of an analogy. *Quarterly Journal of Experimental Psychology*, *28*, 475-482.

Assman, P.F., Nearey, T.M., and Hogan, J.T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, *71*, 975-989.

Baddeley, A.D., and Hitch, G.J. (1977). Recency re-examined. In S. Dornic (Ed.), *Attention and Performance* (Vol. 6, pp. 89-105). New York: Academic Press.

Bjork, R.A., and Whitten, W.B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173-189.

Cole, R.A., Coltheart, M., and Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, *26*, 1-7.

Craik, F.I.M., and Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*, 274-284.

Elman, J.L., and McClelland, J.L. (1984). The interactive model of speech perception. In N.J. Lass (Ed.), *Language and Speech*. New York: Academic Press.

Eukel, B. (1980). A phonotactic basis for word frequency effects: Implications for automatic speech recognition. *Journal of the Acoustical Society of America*, *68*, S33.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales and E. Walker (Eds.), *New Approaches to Language Mechanisms*. Amsterdam: North-Holland.

Forster, K.I. (1979). Levels of processing and the structure of the language processor. In W.E. Cooper and E.C.T. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale, N.J.: Erlbaum.

Greene, R.L. (1986a). Sources of recency effects in free recall. *Psychological Bulletin*, *99*, 221-228.

Greene, R.L. (1986b). A common basis for recency effects in immediate and delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 413-418.

Klatt, D.H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R.A. Cole (Ed.), *Perception and Production of Fluent Speech*. Hillsdale, N.J.: Erlbaum.

Landauer, T.K., and Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Behavior*, *12*, 119-131.

Luce, P.A. (1985). Similarity neighborhoods and word frequency effects in visual word identification: Sources of facilitation and inhibition. Research on Speech Perception Progress Report No. 11, Indiana University, 321-340.

Luce, P.A. (1986). Structural distinctions between high and low frequency words in visual and auditory word recognition. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

Luce, Feustel, T.C., and Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 25, 17-32.

Marslen-Wilson, W.D., and Tyler, L.K. (1980). The temporal structure of spoken language understanding. Cognition, 8, 1-71.

Marslen-Wilson, W.D., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

Mattingly, I.G., Studdert-Kennedy, M., and Magen, H. (1983). Phonological short-term memory preserves phonetic detail. Journal of the Acoustical Society of America, 73, B6.

Nearey, T.M. (1977). Phonetic feature systems for vowels. Doctoral dissertation, University of Alberta. Reprinted by IU Linguistics Club, Bloomington, IN.

Nusbaum, H.C., Pisoni, D.B., and Davis, C.K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. Research on Speech Perception Progress Report No. 10, Indiana University, Bloomington, IN.

Pisoni, D.B., Nusbaum, H.C., and Greene, B.G. (1985). Perception of synthetic speech generated by rule. Proceedings of the IEEE, 73, 1665-1676.

Posner, M.I., Boies, S.J., Eichelman, W.H., and Taylor, R.L. (1969). Retention of visual and name codes of single letters. Journal of Experimental Psychology Monographs, 79, I.

Posner, M.I., and Mitchell, R.F. (1967). Chronometric analysis of classification. Psychological Review, 74, 392-409.

Rabbitt, P. (1968). Channel-capacity, intelligibility and immediate memory. Quarterly Journal of Experimental Psychology, 20, 241-248.

Strange, W., Verbrugge, R.R., Shankweiler, D.P., and Edman, T.R. (1976). Consonant environment specifies vowel identity. Journal of the Acoustical Society of America, 60, 213-224.

Summerfield, Q. (1975). Acoustic and phonetic components of the influence of voice changes and identification times for CVC syllables. Report of Speech Research in Progress, No. 2, 73-98. The Queen's University of Belfast, Belfast, Ireland.

Summerfield, Q., and Haggard, M.P. (1973). Vocal tract normalisation as demonstrated by reaction times. Report on Research in Progress in Speech Perception, 2, 1-12. The Queen's University of Belfast, Belfast, Ireland.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., and Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? Journal of the Acoustical Society of America, 60, 198-212.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses*

Van Summers

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

* A significant portion of the data presented here were reported in my 1985 University of Maryland dissertation. I wish to thank Sigfrid Soli and Betty Tuller for their encouragement and direction in its completion. The dissertation was supported in part by NINCDS Grant NS-13617 to the Haskins Laboratories and by a grant from the Ariel-Benjamin Medical Foundation, directed through the Stuttering Center, Department of Neurology, Baylor College of Medicine. Further data analyses not reported in the dissertation and preparation of the final manuscript were supported by NIH Training Grant NS-07134 to Indiana University in Bloomington. I am grateful to David Pisoni and John Mullennix for comments on an earlier version of this paper.

Abstract

Durations of the vocalic portions of speech are influenced by a large number of linguistic and nonlinguistic factors (e.g., stress, speaking rate, etc.). However, each factor affecting vowel duration may influence articulation in a unique manner. The present study examined the effects of stress and final-consonant voicing on the detailed stucture of articulatory and acoustic patterns in consonant-vowel-consonant (CVC) utterances. Jaw movement trajectories and F1 and F2 patterns were examined for a corpus of utterances differing in stress and final-consonant voicing. Articulator lowering and raising gestures were more rapid, longer in duration, and spatially more extensive for stressed versus unstressed utterances. At the acoustic level, stressed utterances showed more rapid initial F1 transitions and more extreme F1 steady state frequencies than unstressed utterances. In contrast to the results obtained in the analysis of stress, decreases in vowel duration due to voicing did not result in a reduction in the velocity or spatial extent of the articulatory gestures. Similarly, at the acoustic level, the reductions in formant transition slopes and steady state freqencies demonstrated by the shorter, unstressed utterances did not occur for the shorter, voiceless utterances. The results demonstrate that stress-related and voicing-related changes in vowel duration are accomplished by separate and distinct changes in speech production with observable consequences at both the articulatory and acoustic levels.

Effects of Stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses

Speech timing and segmental durations are influenced by multiple linguistic and nonlinguistic factors in fluent speech. Consider a simple consonant-vowel-consonant (CVC) word spoken in a sentence frame. The duration of the vocalic (vowel) portion of this word will be influenced by the inherent or intrinsic duration of the intended vowel, the voicing feature of the following consonant, the speaker's overall speaking rate, the sentential stress pattern, the position of the word within the sentence, and other factors. Previous research suggests that in certain contexts, vowel duration may supply useful perceptual information for many of these factors (Ainsworth, 1972; Denes, 1955; Fry, 1955, 1965; Klatt & Cooper, 1975; Raphael, 1972). However, the fact that segmental durations are influenced by multiple factors makes it difficult to understand how a given pattern of acoustic durations can supply unambiguous perceptual information concerning each factor known to affect the pattern.

An example from the present study may clarify this point. The example deals with the effects of stress and final-consonant voicing on vowel production. Consider the three acoustic waveforms displayed in Figure 1. The waveforms in the top and middle portions of the figure represent the CVC syllable /bab/, produced in a sentence context. The /bab/ token at the top of the figure received primary sentence stress during production. The /bab/ token in the middle of the figure was not stressed. A comparison of the durations of these tokens reflects the general pattern reported in the literature; vowels are generally longer in a stressed context than in an unstressed context (Cooper, Eady, & Mueller, 1985; Fry, 1955; Parmenter & Trevino, 1936). The relationship between stress and vowel duration suggests that vowel duration may provide information about stress with long vowel durations cuing stressed syllables. Now consider the waveform displayed at the bottom of Figure 1. This waveform is based the CVC syllable /bap/. This token received primary stress during production yet it is more similar to the unstressed /bab/ token than to the stressed /bab/ token in terms of vowel duration. The shorter duration of the stressed /bap/ token in comparison to the stressed /bab/ token reflects the influence of final-consonant voicing on vowel duration; vowels followed by voiced final consonants are generally of greater duration than vowels followed by voiceless final consonants (Peterson & Lehiste, 1960; Luce & Charles-Luce, 1985). The unstressed /bab/ and the stressed /bap/ in Figure 1 are similar in duration and are both shorter than the stressed /bab/, but for entirely different reasons. Apparently vowel duration, as an isolated cue, cannot disambiguate durational differences due to stress from differences due to final-consonant voicing.

Insert Figure 1 about here

While ambiguous as an isolated cue to stress or voicing, vowel duration may provide useful information for stress and voicing when used in combination with other perceptual cues. The present study focuses on additional perceptual information contained within the vowel portions of the speech

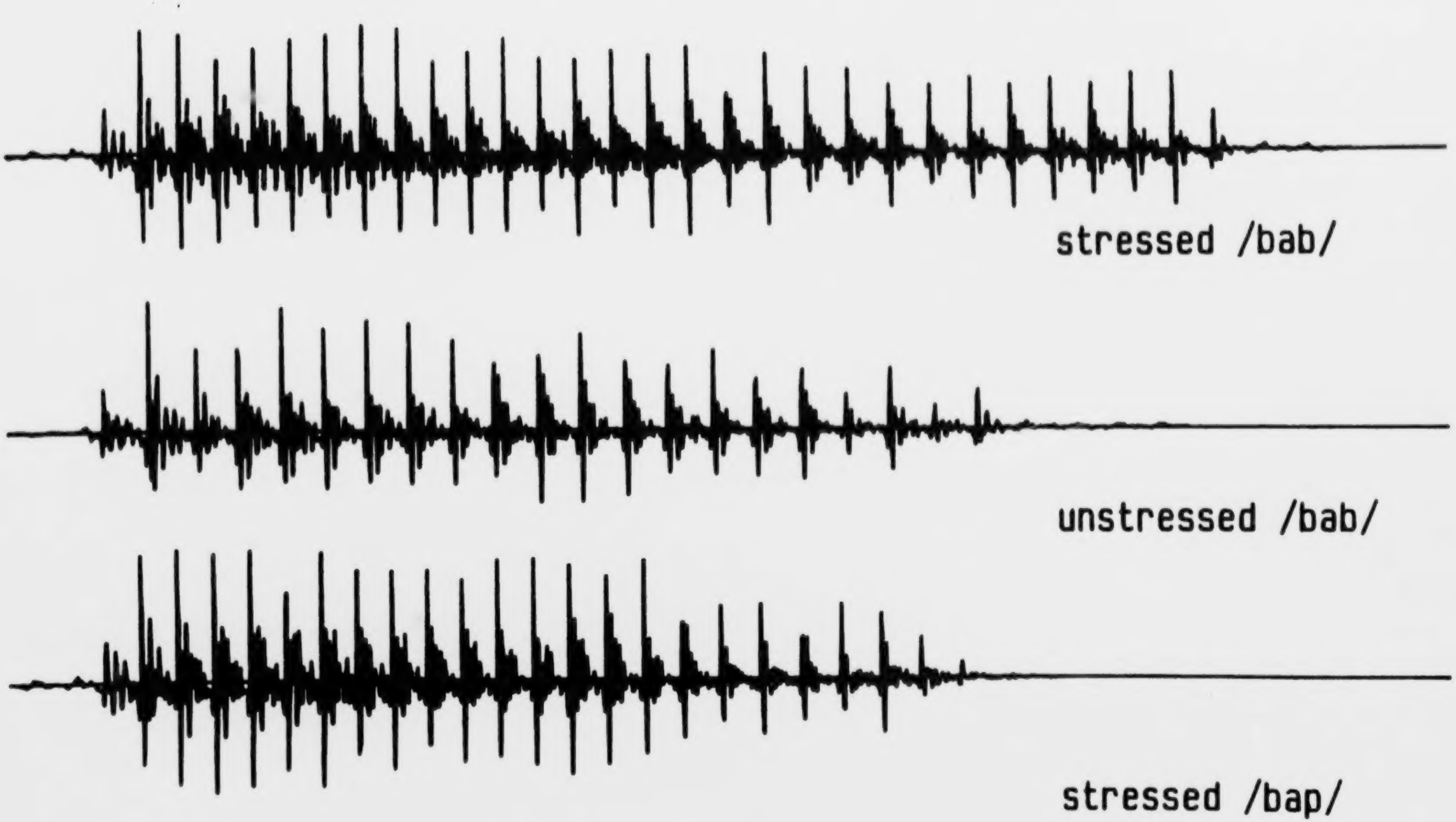


Figure 1. Acoustic waveforms from three utterances differing in stress and final-consonant voicing (speaker: VS).

signal. It may be that each of the factors influencing vowel duration has a unique influence on the detailed acoustic structure of a given vowel (i.e., a specific "signature"). Some support for this hypothesis may be obtained from an examination of the literature concerning the effects of stress and final-consonant voicing on formant patterns.

Regarding stress, it has been reported that in addition to being shorter in duration, unstressed vowels are often reduced or neutralized toward /ʌ/ in formant structure (DeLattre, 1969; Gay, 1978; Harris, 1978; Lindblom, 1963). Thus, a structural change in the formant patterns accompanies the stress-related durational change. While the importance of vowel duration as a cue to stress has been examined by a number of researchers (Fry, 1955, 1965; Morton & Jassam, 1965; Nakatani & Aston, 1978; Westin, Buddenhagen, & Obrecht, 1966), less attention has been given to the potential role of vowel formant structure in providing stress information. As used here, "formant structure" refers to formant properties such as frequencies, transition slopes, transition and steady-state durations, etc. The neutralization of formant frequencies towards /ʌ/ is an example of a change in formant structure which may provide stress information.

Fry (1955, 1965) used two-formant synthetic stimuli to compare syllabic amplitude, duration, fundamental frequency, and formant structure as stress cues in lexical stress pairs (e.g., com'bat and com bat'). By orthogonally varying duration and one of the other variables (e.g., amplitude) while holding the remaining variables constant, Fry ordered the four variables in terms of perceptual importance. His ordering was: fundamental frequency, duration, amplitude, and formant structure. Unfortunately, Fry's experimental techniques were extremely primitive. Fry (1965) varied the steady-state frequency of synthetic vowels in order to examine how vowel reduction affects stress judgments. The naturalness of these stimuli and the extent to which changing steady-state target frequency approximates normal vowel reduction are both open to question. Thus Fry's (1965) experiments may be a poor test of formant structure as a stress cue. Summers (1981) reexamined this issue using natural productions of lexical stress pairs. Using computer editing, he increased the durations of unstressed vowels to compare the importance of duration versus structure as cues to stress. His results showed, in contrast to Fry's (1965) earlier findings, that vowel structure outweighed vowel duration as a cue to stress in many cases. Thus listeners may rely more on vowel structure for stress information than Fry's (1955,1965) results would lead one to expect.

Reduction of formant frequencies towards /ʌ/-like values does not appear to accompany voicing-related reductions in vowel duration. However, consistent voicing-related changes in formant structure have been reported in the literature. Soli (1982) carried acoustic analyses of productions of /jus/ and /juz/ by two speakers at two speaking rates and reported that /jus/ tokens had proportionally longer initial transitions than /juz/ tokens which had proportionally longer steady-state regions.

Perceptual experiments have demonstrated the importance of formant structure in cuing final-consonant voicing. Fitch (1981) demonstrated that when vowel duration and closure duration were held constant, a vowel containing a longer initial transition was more likely to be heard as preceding a voiceless stop while a vowel containing a longer steady-state region was more likely to be heard as preceding a voiced stop. Thus Fitch's (1981) results with stop consonants and Soli's (1982) findings with fricatives suggest the same formant structure component as a cue to final voicing. With the exception of these two studies, previous research has devoted more attention to vowel duration than formant structure as a cue to final voicing (Derr & Massaro, 1980; Luce & Charles-Luce, 1985; Raphael, 1972). Other research has focused on the ratio of vowel duration to fricative or closure duration as a cue to final-consonant voicing (Denes, 1955; Port, 1981; Port & Dalby, 1982).

The present research examines changes in the detailed articulatory and acoustic structure of vowels which accompany stress-related and voicing-related changes in vowel duration. Articulatory movement data and formant data were collected for a set of CVC utterances varying in stress and final-consonant voicing. The articulatory data were obtained through an optical tracking system which monitored the movement of light-emitting diodes (LED's) attached to subjects' lips and jaw during production. The formant data were obtained by simultaneously recording the speech signal. Specialized hardware and software allowed for the storage and analysis of the articulatory and acoustic data in a time-locked fashion.

The goal of this research was to describe how changes in stress and final-consonant voicing affect articulatory movement patterns and formant patterns. Consistent stress-related and voicing-related changes in movement patterns and formant patterns may convey perceptual information for stress and voicing. Previous research has already identified some of the ways in which stress may influence articulatory and acoustic characteristics of vowels. This previous work allows one to generate several predictions about the effects of stress on movement patterns and formant patterns in the present study. For example, in comparison to stressed utterances, unstressed utterances should display reductions in the maximum displacement of the lips and jaw (Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Kent & Netsell, 1971) and concommitant reductions in formant frequencies toward more central (/ʌ/-like) values (Lindblom, 1963). In reducing maximum displacement, the entire articulatory movement pattern is restructured. Similarly, in reducing steady-state formant frequencies, the formant trajectories are restructured. The present focus is on describing this restructuring at both the articulatory and acoustic levels. Changes in the entire formant pattern rather than changes in steady-state target frequencies may provide the relevant perceptual information for assignment of stress. By collecting productions of utterances which vary in stress and final-consonant voicing, and by examining the articulatory movement patterns and formant patterns associated with vowels from these utterances, it should be possible to identify aspects of articulatory and acoustic structure which vary across utterances differing in stress and final-consonant voicing. Characteristics of the articulatory and acoustic patterns which show clear differences across stress levels and across voicing conditions may provide the listener with perceptual information that is used in conjunction with vowel duration to specify stress and final-consonant voicing.

A further goal of this research was to map stress-related and voicing-related changes in the movement patterns to changes in the formant patterns and vice versa. The relations we examined between the articulatory and acoustic data were fairly straightforward. The main articulatory and acoustic data to be reported are jaw movement data and F1 data. The test utterances were CVC's containing the vowels /a/ or /æ/. These vowels are produced with the tongue low in the oral cavity which produces a high first formant frequency. In the present study we assumed that jaw height and tongue height bear the same relationship to F1 frequency. Thus, low jaw positions are expected to correlate with high F1 frequencies. Slopes of the jaw movement gestures will be compared with slopes of F1 transitions. For example, a rapid initial jaw-lowering gesture is expected to correlate with a steep F1 initial transition. Finally, durations of various components of the articulatory gestures will be compared with durations of components of the F1 trajectory. For example, a long jaw-raising duration at the end of a vowel is expected to correlate with a long F1 final transition.

Method

Stimuli. The utterances examined were CVC syllables with initial consonant =/b/, vowel =/a/ or /æ/, and final consonant =/b/, /p/, /v/, or /f/. Thus the stimulus set contained final voiced and voiceless stops and final voiced and voiceless fricatives. The vowels and final consonants chosen involve significant jaw and lip movement in their production.

Procedure. Three male native English speakers (BW, VS, and EB) produced multiple repetitions of each test stimulus (e.g., /bab/) embedded in the sentence frame: "I'll be at the bus CVC up the street." Speakers were instructed to produce each sentence with contrastive stress on the target utterance or with stress on the word immediately preceding it, leaving the target utterance unstressed. Speakers produced 10 repetitions of each utterance at each stress level. A total of 160 utterances were produced by each speaker (2 stress levels x 2 medial vowels x 4 final consonants x 10 repetitions). Tokens were repeated in instances in which the speaker or either of two judges believed that a mispronunciation of the desired utterance had occured.

The Haskins Laboratory's Selspot optical tracking system was used to obtain jaw and lower lip position data during utterance production. During production by each speaker, the Selspot camera tracked the movements of infrared LED's placed at four locations: the midline of the vermilion border of the upper and lower lips, the tip of the nose, and the the point of the jaw. Each speaker was seated 21 inches in front of the Selspot camera with his face parallel to the camera's focal plane. A microphone was positioned to one side of the speaker's face so that it did not occlude any of the LED's. Prior to stimulus production, the position of each LED was recorded with the jaw and lips in a normal resting position. As the utterances were being spoken, the speech waveform, position data from each LED, and a timing signal produced once every second were recorded on separate channels of an FM tape recorder.

The Haskins Laboratory's Physiological Signal Processing (PSP) software system (Gulisano, 1982) was used to sample the movement records and associated speech waveforms into a PDP 11/45 computer and to store these data in a time-locked fashion. The LED movement data were sampled by the computer at a rate of 200 Hz and the speech waveform data were sampled at 10,000 Hz. The LED movement data were numerically smoothed using a 25 ms triangular window. Synchronization of the speech and movement data was maintained to within 2.5 ms of accuracy with the PSP software.

Measurement of total vowel duration. Total vowel duration of each utterance was determined from visual inspection of a CRT display of its acoustic waveform. For utterances containing final stop consonants, total vowel duration was defined as the period from voicing onset following release of the initial /b/ to onset of final closure. For utterances containing final fricatives, total vowel duration was defined as the period from voicing onset to frication onset. Final closure, for target utterances containing final stops, and frication onset, for utterances containing final fricatives, will henceforth be referred to as "vowel offset". Voicing onset following release of the initial /b/ will be referred to as "vowel onset".

Articulatory analyses. Vertical movements of the jaw and lower lip were examined for the vowel portion of each utterance. The jaw movement data will be described in this report. The lower lip data are not included because they did not differ in any important respect from the jaw results. The 200 Hz sampling rate provided jaw position data at 5 ms intervals.

Figure 2 displays jaw movement traces for the three utterances shown in Figure 1. The figure displays clear differences in jaw movement between utterances differing in stress and final voicing. These differences will be examined at length in later sections. The figure is provided here to help clarify the segmentation process used in analyzing the articulatory data, which will now be described.

Insert Figure 2 about here

Jaw movement traces from each utterance were segmented into three sections corresponding to an initial jaw lowering portion, a steady-state portion between lowering offset and raising onset, and a final raising portion. The following segmentation rule was used to divide each jaw trace into these three components. First, jaw position at vowel onset and position at maximum lowering were determined. The total change in jaw position from vowel onset to maximum lowering was then calculated. A cutoff value was then established which corresponded to 80% of the change in position from vowel onset to maximum lowering (position at vowel onset + .8(position at maximum lowering - position at vowel onset). The first point in the movement trace with a value below this cutoff was defined as the first point of the steady-state region. Similarly, the last point in the trace with a value below this cutoff was defined as the final point in the steady-state region. The use of this rule provided an objective method of dividing each jaw movement trace into an initial lowering portion, a steady-state portion, and a final raising portion. As will be seen, the same segmentation strategy was applied to the F1 data, dividing each formant pattern into an initial rising transition, a steady-state region, and a final falling transition.

Formant analyses. Linear predictive coding (LPC) analysis was used to examine the spectral-temporal structure of F1 and F2 for the vowel portion of each utterance. Linear prediction coefficients were calculated every 5 ms using the autocorrelation method. Thus the spectral analysis matched the articulatory analysis in using a 5 ms interval between samples. Fourteen LPC coefficients and a 20 ms Hamming window were used in the LPC analyses. A peak-picking algorithm was applied to the LPC spectra to estimate formant frequencies (ILS version 5.0, Signal Technologies Corporation, 1985). Figure 3 displays F1 frequency traces for the utterances shown in Figure 1. The figure is provided to clarify the following description of the segmentation of the F1 frequency traces.

Insert Figure 3 about here

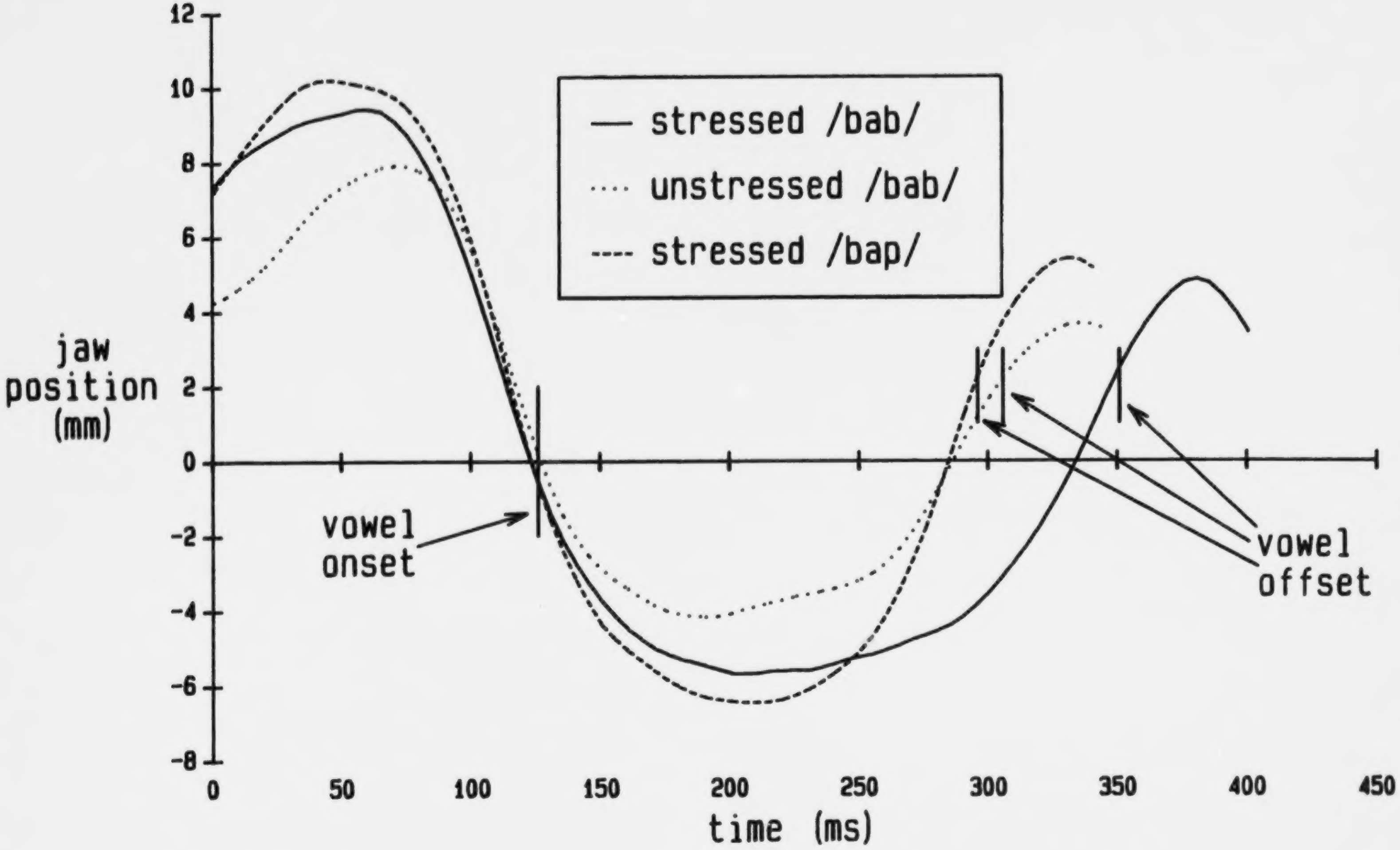


Figure 2. Jaw trajectories for utterances shown in Figure 1. Jaw resting position equals 0 mm.

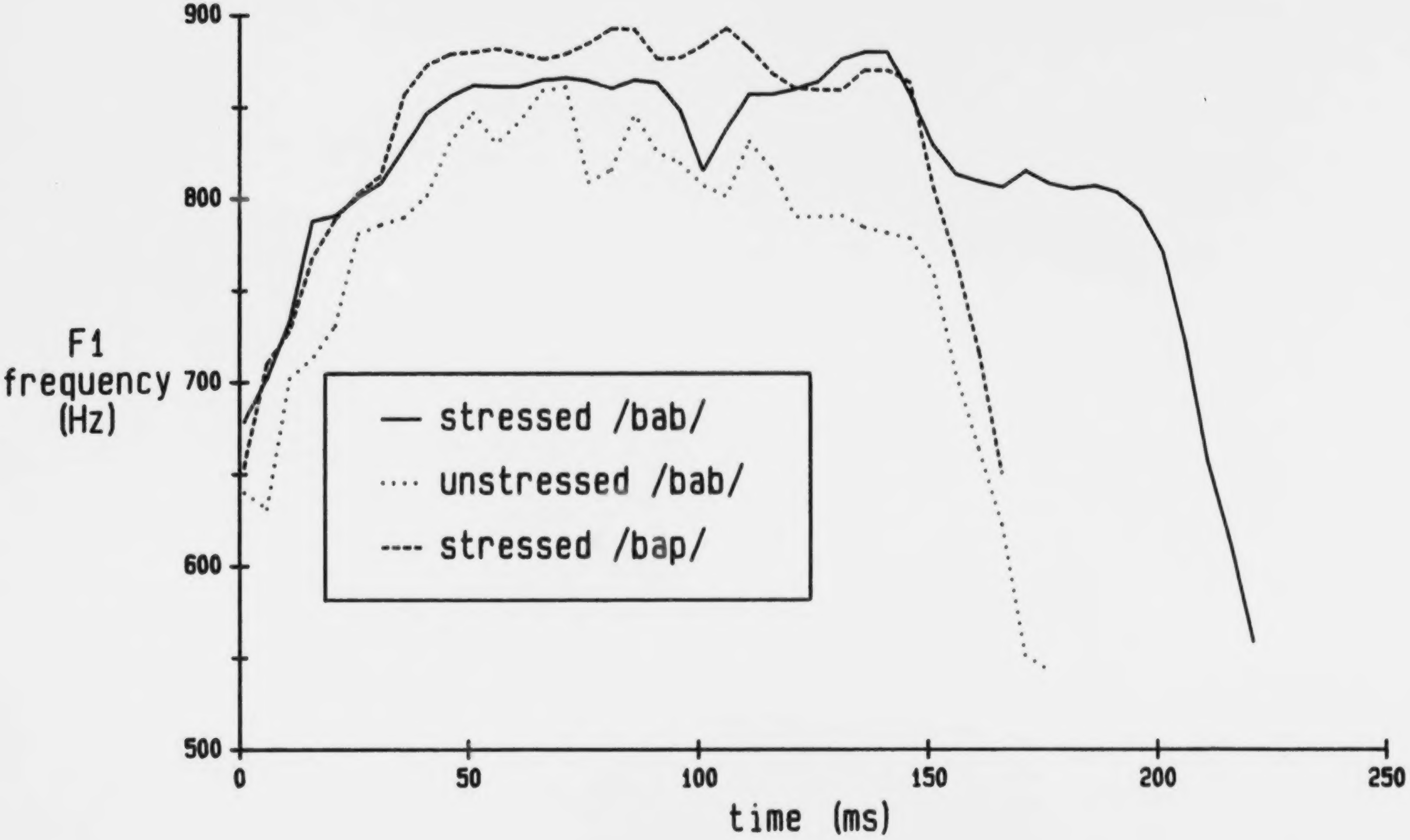


Figure 3. First formant trajectories for utterances shown in Figure 1.

As already mentioned, the F1 data from each utterance were segmented into three sections: a rising initial transition, a steady-state region, and a falling final transition. The segmentation rule used to divide each F1 trace was the same one used in segmenting the jaw movement traces. First, the total change in F1 frequency from vowel onset to peak F1 frequency was calculated. A cutoff value was then established which corresponded to 80% of the change in frequency from vowel onset to peak F1 frequency. The first point in the F1 trace with a frequency greater than this cutoff value was defined as the first point of the steady-state region. The last point in the trace with a value greater than this cutoff was defined as the final point in the steady-state region.

By applying the same segmentation strategy to both the jaw movement data and the F1 data, each set of data was divided into initial transitional regions, medial steady-state regions, and final transitional regions. The durations of these three regions were examined for each movement trace and each F1 trace. The slopes of the transitional regions were also examined for each trace. Initial transition slopes were calculated by dividing the change in position (or frequency) from vowel onset to steady-state (the 80% cutoff value) by the duration of the initial transition. Final transition slopes were calculated in an analogous fashion. Finally, jaw positions and F1 frequencies at vowel onset, at vowel offset, and at the cutoff value used to define the steady-state region, were examined. Altogether, a total of eight descriptive properties were examined for each jaw movement trace and for each F1 trace. Regression analyses, described in the following section, were used to determine the influence of stress and final-consonant voicing on each of these jaw movement and F1 properties.

Stress and final-consonant voicing influences on jaw movement and F1 structure are the main foci of this study. One additional acoustic measurement included in the analysis was F2 peak frequency. F2 structure was not examined in detail in the present study for two reasons. First, given that the articulatory data are vertical (up-down) position data, there was little expectation that these data would bear a close relationship to F2 which is generally associated with the horizontal (front-back) position of the tongue. Secondly, a number of the F2 trajectories did not contain initial formant transitions and therefore could not be segmented into transitions and steady-state regions using the segmentation strategy described above. The analysis of F2 peak frequency in the present study was included in order to determine the extent to which vowel neutralization (movement of formant frequencies towards /ʌ/-like values) in F1 was independent of or correlated with neutralization in F2.

Statistical Analyses. The influence of stress and final-consonant voicing on various articulatory and acoustic properties of the test utterances was examined using multiple regression techniques. Each utterance was coded in terms of stress, final-consonant voicing, final-consonant manner , and medial vowel. Since each of these variables is dichotomous (e.g., stressed, unstressed), each could be described in a single vector containing the values 1 and -1 (i.e., contrast coding). Interactions were also coded into vectors containing the values 1 and -1 by multiplying the values from the main effect vectors involved in the interaction. Articulatory and acoustic properties of the test utterances (e.g., F1 steady-state frequency, jaw steady-state position) were then used as dependent variables in regression analyses with stress, final-consonant voicing, etc., serving as independent or predictor variables. The question of whether a given predictor variable (e.g., stress) was significantly related to a given dependent variable (e.g., F1 steady-state frequency) was addressed by examining whether the vector coding that predictor made a significant contribution to the prediction equation. This was determined by testing the regression coefficient (the b weight) of each vector for a significant deviation from zero. Separate analyses were carried out for each of the three speakers. Each regression analyses used in this study included 15 predictor variables (the main effects of stress, final-consonant voicing, vowel, and final-consonant manner, and all possible interactions among these factors). Three of these 15 factors were not examined or tested for significance since they did not involve stress or final-consonant voicing (the factors not tested were the vowel and final-consonant manner main effects and the vowel x manner interaction). Therefore, a total of 12 predictor variables were tested for significance in each analysis. The large number of analyses carried out in the present study and the large number of significance tests carried it out within each analysis made it necessary to adopt a fairly stringent alpha level for testing each individual predictor variable. A *p* value of .004 was used as the critical value in all tests of individual predictor variables. The probability of a Type 1 error when 12 significance tests are carried out at .004 is approximately .05. This is the overall probability of a Type 1 error for each analysis of a given acoustic or articulatory property in the present study.

Regression analyses were used in the present study rather than the more common analysis of variance approach in order to deal with cases of missing data. Three utterances were unavailable due to mistakes in the original computer-sampling of the utterances. In addition, for each speaker there were several utterances for which reliable F1 tracks could not be obtained. These utterances were excluded from all analyses including the analyses of total vowel duration and of jaw movement properties. The statistical procedures involved in the repeated-measures analysis of variance that would be appropriate to the present experiment are not easily modified for use in an experiment in which each cell does not contain the same number of data points (i.e., utterances). The multiple regression techniques used here do not require equal cell N's and are therefore more appropriate for the present data. The analysis of variance and regression approaches are statistically identical in cases of equal cell N's. Of the 160 utterances produced by each speaker, 155 were available for the analyses of BW's productions and 156 were available for EB and VS. At least 8 of the original 10 utterances were available in each utterance category for each speaker.

The unequal cell N's in the present study make arithmetic means only meaningful at the individual cell level. Above the cell level, the appropriate comparison is between least squares estimated means based on the regression equation (at the individual cell level least squares estimated means and arithmetic means are identical). All references to means in the following sections refer to least squares estimated means.

Results I

Influence of Stress and Final-Consonant Voicing on Total Vowel Duration

This section will describe stress and final-consonant voicing effects on total vowel duration. As already noted, previous research has demonstrated that each of these factors significantly influences vowel duration. The expected stress-related and voicing-related differences in vowel duration must be demonstrated in the present data prior to examining changes in the articulatory and formant data which accompany these changes in duration.

Mean total vowel durations for stressed utterances and unstressed utterances are presented for each speaker in the lefthand panel of Figure 4. For each speaker, stress had a significant influence on the total vowel duration. Stressed utterances (S+ utterances hereafter) were longer in duration than unstressed utterances (S- utterances) ($p < .0001$ for each speaker). These results are in agreement with previous studies reporting stress-related differences in vowel duration (Cooper, Eady, & Mueller, 1985; Lieberman, 1960; Oller, 1973; Parmenter & Trevino, 1936; Summers, 1981).

Insert Figure 4 about here

Mean total vowel durations for utterances containing voiced final consonants (V+ utterances hereafter) and for utterances containing voiceless final consonants (V- utterances) are presented in the righthand panel of Figure 4. V+ utterances were significantly longer in duration than V- utterances for each speaker (ps < .0001). Thus, the expected voicing-related differences in vowel duration were also present in these data (Denes, 1955; House & Fairbanks, 1953; Luce & Charles-Luce, 1985; Mack; 1982; Peterson & Lehiste, 1960).

For speaker EB, the stress and voicing main effects were mediated by a significant stress x voicing interaction ($p < .0001$). The stress x voicing interactions for speakers VS and BW fell short of the .004 significance level ($p = .0101$ (VS) and $p = .0089$ (BW)). However, the nature of the stress x voicing interaction was similar across speakers. In each case, while S+ utterances showed greater vowel durations than S- utterances regardless of final-consonant voicing, stress had a greater effect on duration for utterances containing voiced final consonants than voiceless final consonants. For EB, only a very small change in vowel duration (approximately 2.5 ms) was observed across stress conditions for V- utterances. The stress x voicing interaction can also be described in terms of the final-consonant voicing effect. While V+ utterances showed longer vowel durations than V- utterances within each stress condition, voicing had a greater influence on duration for stressed utterances than unstressed utterances for each speaker.

One way of describing these interactions is to assume that a change in stress did not influence total vowel duration as much for the durationally shorter voiceless utterances as for the longer voiced utterances. Alternatively, final-consonant voicing did not influence the duration of the shorter unstressed utterances as much as it did for the longer stressed utterances. These results would be expected given Klatt's (1973, 1975, 1976)

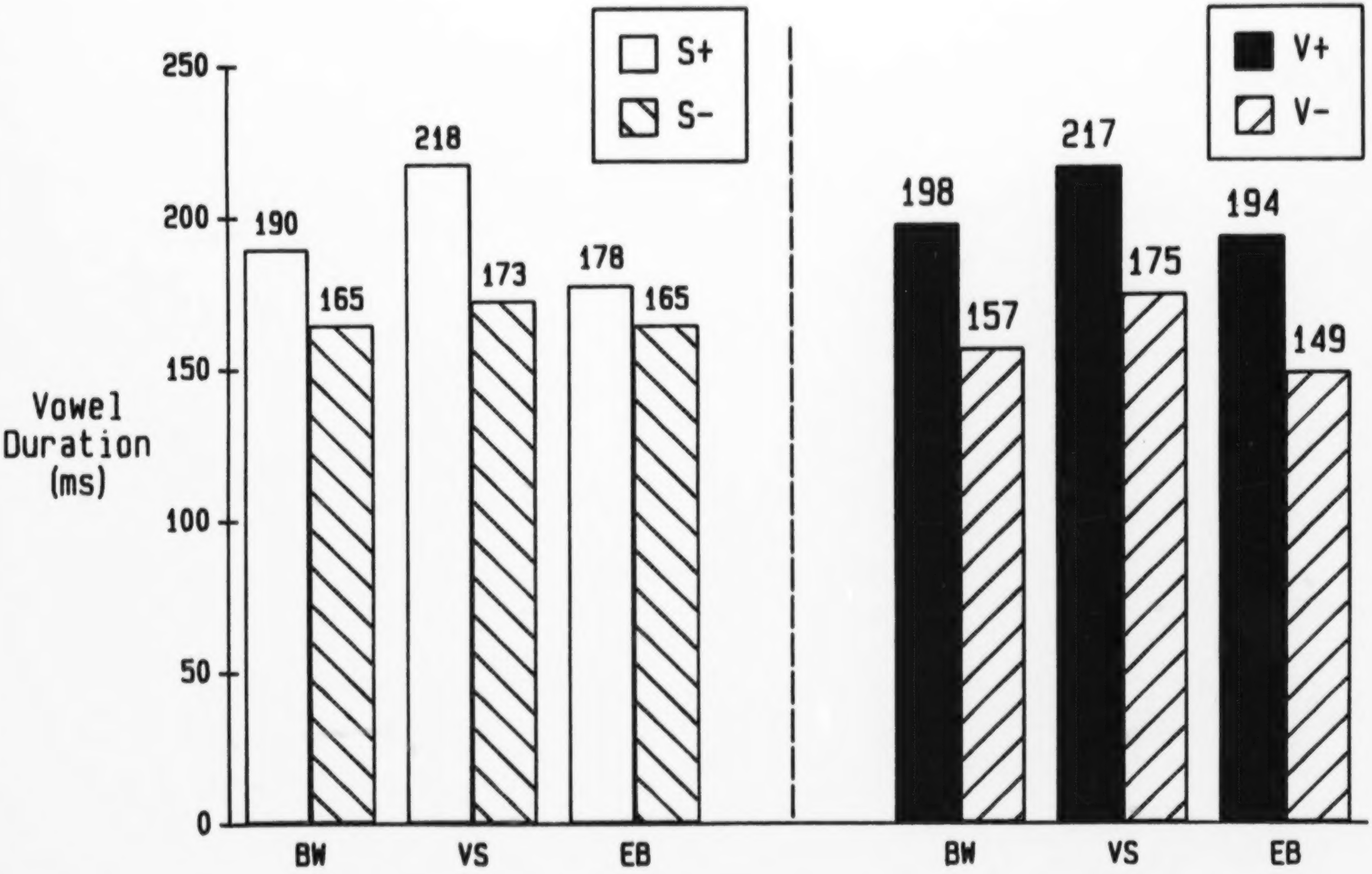


Figure 4. Mean vowel durations for stressed (S+) versus unstressed (S-) utterances (lefthand panel) and for utterances containing voiced (V+) versus voiceless (V-) final consonants (righthand panel).

proposal on the "incompressibility" of segment durations beyond some minimum value. The present data support incompressibility if it is assumed that destressing (or devoicing) did not decrease the duration of voiceless (or unstressed) utterances by a larger amount due to a limit on minimum vowel duration. Klatt has suggested that this minimum duration "reflects a minimum time of execution of the required articulatory program" (Klatt, 1973, p. 1103).

For speaker BW, the stress effect was also mediated by a significant stress x final-consonant manner interaction ($p < .0001$). The influence of stress on vowel duration was greater for utterances containing final fricatives than final stops. This pattern is again in accord with Klatt's incompressibility proposal. Once again the durationally shorter utterances (in this case, the utterances containing final stop consonants) showed less durational change across stress conditions than the longer (fricative) utterances.

Taken together, the results demonstrate that the expected stress-related and voicing-related differences in total vowel duration are present in these data. The following sections examine changes at the articulatory level and changes in F1 and F2 which accompany stress and voicing-related changes in vowel duration.

Results II

Influence of Stress and Final-Consonant Voicing on Jaw Movement

This section examines the effects of stress and final-consonant voicing on jaw movement patterns during vowel articulation. Eight descriptors of each articulatory gesture were examined. The initial jaw-lowering portion of the gesture was examined in terms of jaw position at voicing onset, lowering gesture slope, lowering gesture duration, and position at lowering offset ("steady-state" position as defined in the Method section). The steady-state portion of the gesture following lowering offset and preceding raising onset was examined in terms of jaw position and steady-state duration. The final, raising portion of the gesture was examined in terms of jaw position at raising onset (steady-state position), raising gesture slope, raising gesture duration, and position at vowel offset. Mean values for stressed versus unstressed utterances for each of these articulatory variables are listed for each speaker in Table I. Mean values for V+ versus V- utterances are listed in the righthand portion of the table.

Insert Table I, Figure 5, and Figure 6 about here

The mean values listed in Table I for S+ and S- utterances (with the exception of lowering and raising gesture slopes) are represented graphically in the three panels of Figure 5. Each speaker's data appear in a separate panel. Within each panel of the figure, mean values for stressed versus unstressed utterances are represented by lines connecting four points (these four points are labelled for the line representing S+ utterances in the upper panel of Figure 5). The first point represents jaw position at vowel onset. The second and third points represent jaw steady-state position (the 80%

Table I. Least squares estimated means on articulatory movement variables for S+ versus S- utterances and V+ versus V- utterances.

jaw property	Speaker	stressed utterances	unstressed utterances	stressed - unstressed	voiced utterances	voiceless utterances	voiced - voiceless
position at vowel onset (mm) (0 mm = resting)	VS	-0.95	-1.03	0.07	-0.97	-1.01	0.04
	BW	-1.42	-1.67	0.25	-1.59	-1.49	-0.10
	EB	-3.22	-3.43	0.20	-3.16	-3.48	0.32
lowering slope (Hz/sec)	VS	-10.56	-8.25	-2.31 ***	-9.14	-9.67	.53
	BW	-6.17	-4.92	-1.25 ***	-5.16	-5.93	.77 **
	EB	-13.36	-8.65	-4.71 ***	-10.26	-11.75	1.49 ***
lowering duration (ms)	VS	42	35	7 ***	38	39	-1
	BW	40	38	2	37	41	-4
	EB	38	35	3 **	36	36	0
steady-state position (mm)	VS	-5.39	-3.89	-1.50 ***	-4.45	-4.83	0.38 ***
	BW	-3.87	-3.51	-0.37 ***	-3.49	-3.89	0.40 ***
	EB	-8.27	-6.47	-1.80 ***	-6.99	-7.75	0.76 ***
steady-state duration (ms)	VS	94	72	22 ***	92	73	19 ***
	BW	83	64	19 ***	84	62	22 ***
	EB	75	55	20 ***	74	55	19 ***
raising slope (Hz/sec)	VS	11.31	10.10	1.20 ***	8.63	12.78	-4.15 ***
	BW	10.75	9.24	1.51 ***	9.40	10.59	-1.19 **
	EB	16.60	11.10	5.50 ***	11.24	16.46	-5.22 ***
raising duration (ms)	VS	83	67	16 ***	87	62	25 ***
	BW	68	63	5	76	54	22 ***
	EB	65	75	-10 **	82	58	24 ***
position at vowel offset (mm)	VS	3.38	2.59	0.80 ***	2.98	2.99	-0.01
	BW	2.84	1.97	0.87 ***	3.14	1.67	1.46 ***
	EB	1.57	1.13	0.43	1.46	1.24	0.22

*	$p < .004$
**	$p < .001$
***	$p < .0001$

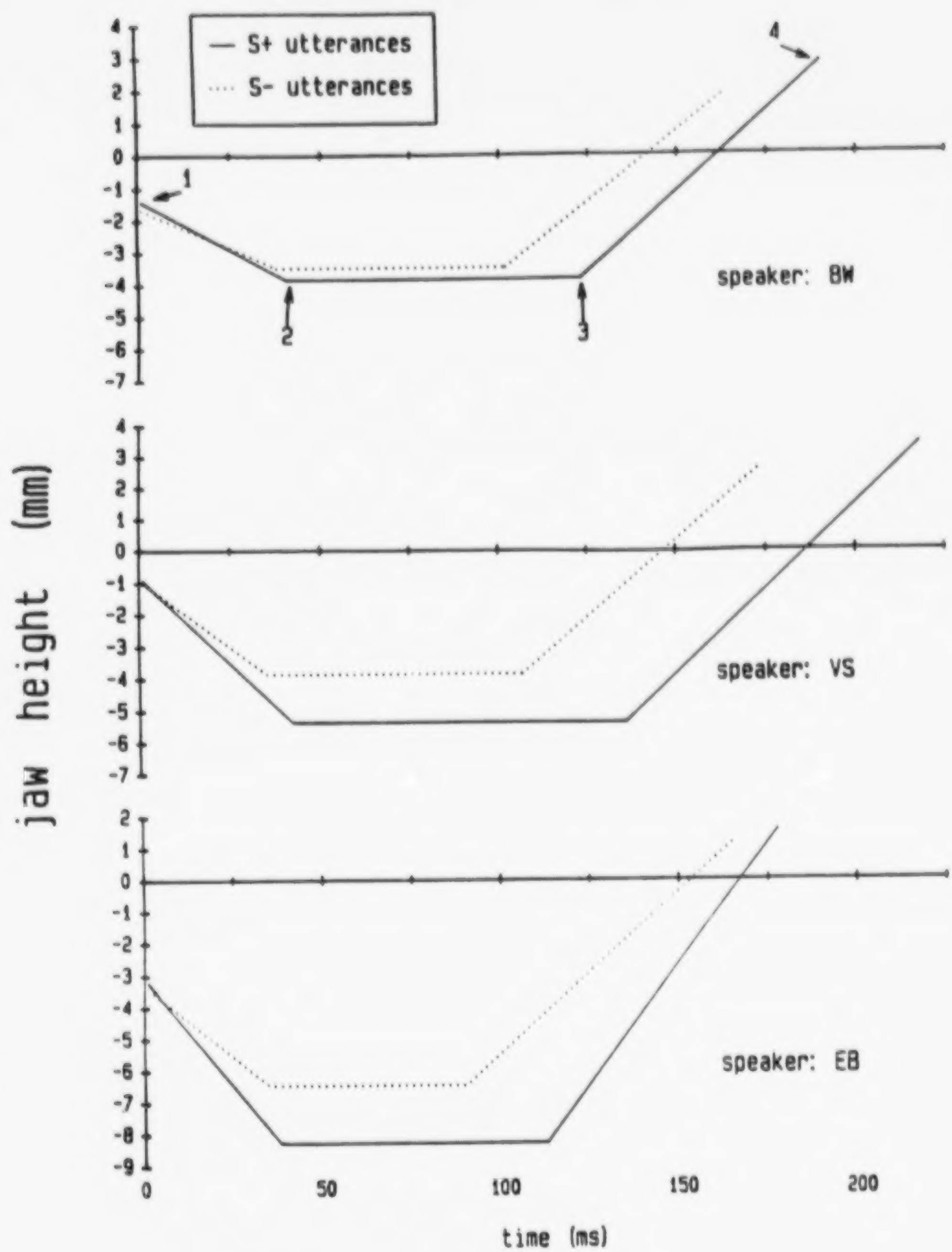


Figure 5. Jaw position plots for stressed (S+) versus unstressed (S-) utterances. Based on mean positions and mean durations listed in Table I.

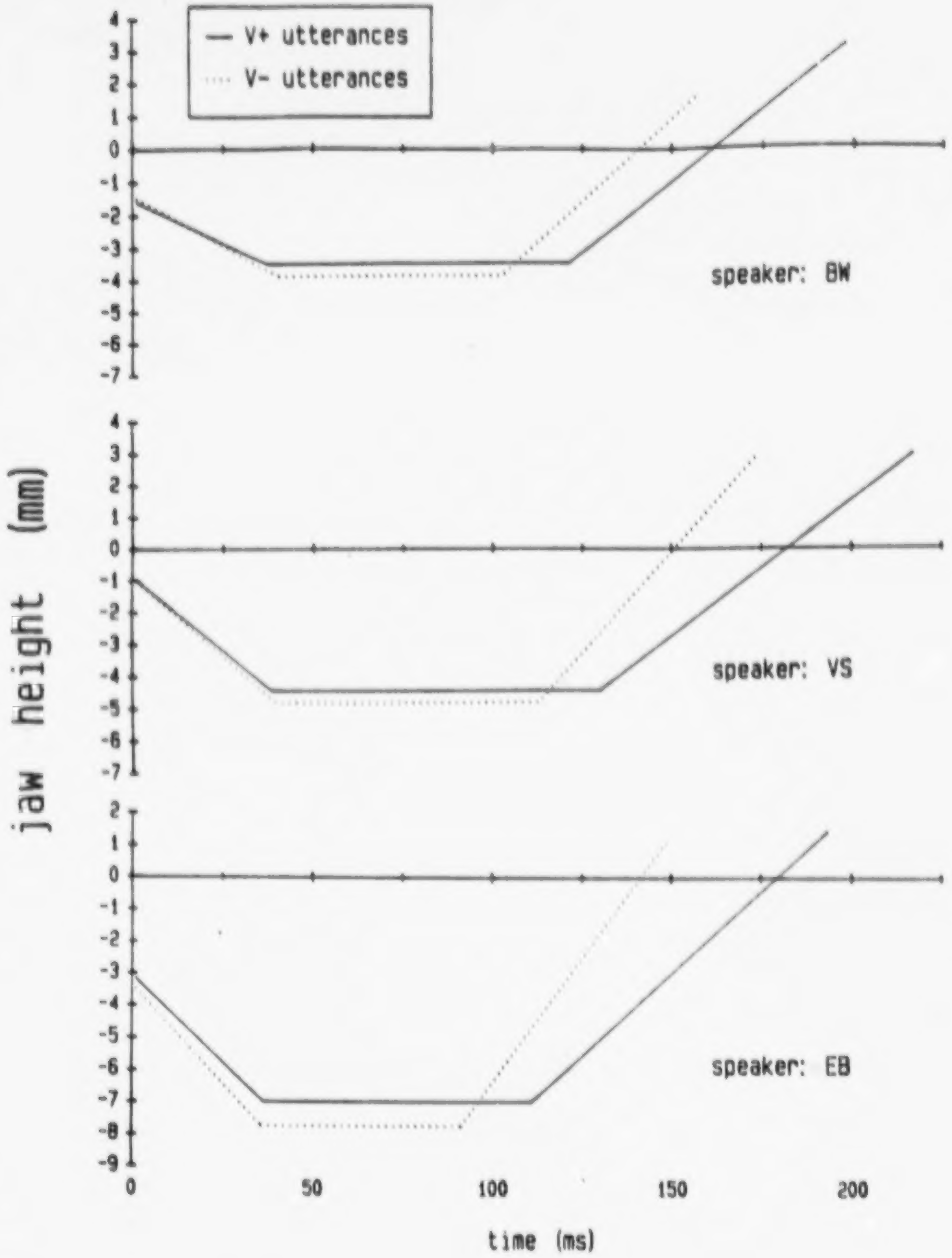


Figure 6. Jaw position plots for utterances containing voiced (V+) versus voiceless (V-) final consonants. Based on mean positions and mean durations listed in Table I.

cutoff value). The final point represents position at vowel offset. The distance along the x axis between points 1 and 2 represents jaw lowering duration. The x-distance between points 2 and 3 represents jaw steady-state duration. The x-distance between points 3 and 4 represent jaw raising duration. The mean jaw positions and mean durations for V+ and V- utterances listed in Table I are shown in the panels of Figure 6 using the same format as Figure 5. The mean slopes of the initial lowering gestures and final raising gestures listed in Table I were not used in constructing Figures 5 and 6. However, the slopes of the transitions in these figures are representative of these mean slopes in most instances.

Influence of Stress and Voicing on Jaw Lowering

Jaw Position at Vowel Onset

For each speaker, mean jaw positions at vowel onset were slightly higher for S+ utterances than S- utterances. However, differences in jaw position across stress levels were not significant for any of the three speakers. Similarly, final-consonant voicing did not significantly influence jaw position at vowel onset for any of the speakers.

Jaw Lowering Slope

For each speaker, lowering gesture slopes were significantly steeper in S+ utterances than in S- utterances ($p <$ .0001 for each speaker). For BW, this main effect was mediated by a significant stress x voicing x vowel interaction ($p =$ .0037). This interaction reflected a smaller stress effect for /a/ utterances containing voiceless final consonants than for other utterances. The stress effect was consistent in direction (i.e., S+ utterances showed steeper lowering slopes than S- utterances) within each voicing x vowel condition. The steeper slopes demonstrated by S+ utterances for each speaker suggest that these utterances are produced with more rapid jaw lowering than occurs for S- utterances.

Final-consonant voicing also influenced jaw lowering slopes. For BW and EB, V- utterances demonstrated significantly steeper jaw lowering slopes than V+ utterances ($p =$ .0004 (BW);$p =$ .0001 (EB)). The main effect of voicing was nonsignificant for VS ($p =$ .0972) although V- utterances again showed steeper slopes than V+ utterances. A significant stress x voicing x vowel interaction mediated the voicing main effect for BW (the interaction mentioned in the previous paragraph in conjunction with the stress main effect). For this speaker, V- utterances did not have steeper lowering slopes than V+ utterances in stressed utterances containing /a/. The change in slope across voicing conditions was larger in each of the other stress x vowel conditions, all of which demonstrated steeper slopes for V- utterances, than in this one case where V+ utterances demonstrated steeper slopes. Although the results concerning the effect of final-consonant voicing on lowering gesture slopes are not consistent, the data suggest that V- utterances tend to demonstrate steeper lowering slopes than V+ utterances, suggesting that more rapid jaw lowering may be associated with voiceless final consonants.

Jaw Lowering Duration

In addition to displaying steeper slopes, lowering gestures for S+ utterances were also durationally longer than for S- utterances. For EB and VS, S+ utterances showed significantly greater lowering durations than S- utterances ($p =$ .001 (EB); $p <$ .0001 (VS)). S+ utterances also demonstrated longer durations for BW although the stress effect was not statistically significant (p= .0441).

Final-consonant voicing did not significantly influence jaw lowering duration for any speaker. Thus it appears that final-consonant voicing influences total vowel duration without effecting the duration of the initial jaw-lowering gesture.

Jaw Steady-state Position (position at lowering offset)

Stress had a clear and consistent influence on jaw position at lowering offset. S+ utterances showed significantly lower jaw positions at lowering offset than S- utterances for each speaker (p< .0001 for each speaker).

Final-consonant voicing also had a significant effect on jaw position at lowering offset. For each speaker, V- utterances demonstrated lower offset positions than V+ utterances (p< .0001 for each speaker).

To summarize the data presented thus far, stress had an influence on the velocity (slope), duration, and extent of jaw lowering gestures. Compared to unstressed utterances, stressed utterances were produced with more rapid lowering gestures of greater duration. The increase in velocity and duration produced more extreme (lower) articulatory positions at the termination of lowering for each speaker. Stress effects on lowering gesture slope and position at lowering offset were more consistent than on lowering gesture duration for speaker BW.

Final-consonant voicing had less influence than stress on jaw lowering. Voicing did influence jaw-lowering slopes with V- utterances having slightly steeper lowering slopes than V+ utterances. In addition, the jaw reached a more extreme (lower) position at lowering offset in V- utterances than in V+ utterances. With no consistent voicing influence on jaw position at vowel onset or on duration of the jaw lowering gesture, it appears that the slight increase in the slope of the lowering gesture seen for V- utterances allowed these utterances to achieve lower jaw positions at lowering offset. Voicing-related differences in jaw lowering slope and position at lowering offset were generally not as large as the stress-related differences in these same variables.

Influence of Stress and Voicing on Jaw Steady-State Region

As described above, S+ utterances displayed consistently lower steady-state positions than S- utterances. The duration of the steady-state region was also significantly influenced by stress; S+ utterances showed longer steady-state regions than S- utterances for each speaker (ps < .0001). The stress main effects were mediated by significant stress x voicing interactions for BW and EB (p= .0008 (BW);p< .0001 (EB). For these speakers, stress-related differences in jaw steady-state duration were greater for V+ utterances than V- utterances. Thus, jaw steady-state position and steady-state duration were both significantly influenced by stress; S+ utterances displayed more extreme (lower) jaw steady-state positions and greater steady-state durations than S- utterances.

As already described, final-consonant voicing influenced jaw steady-state position; V- utterances displayed lower steady-state positions than V+ utterances. The duration of the steady-state region was also significantly influenced by voicing; V+ utterances showed longer steady-state regions than V- utterances for each speaker (ps < .0001). As mentioned above, significant stress x voicing interactions mediated the stress and voicing main effects for BW and EB. For these two speakers, voicing influenced duration more in S+ utterances than in S- utterances. Significant voicing x manner interactions occured in the analyses of jaw steady-state duration for BW and EB as well (p= .0039 (BW);p< .0001 (EB)). The change in duration across voicing conditions was greater in utterances containing final fricatives than final stops for these two speakers.

The present results suggest that while V- utterances demonstrate more extreme steady-state positions than V+ utterances, steady-state positions are maintained for briefer periods in V- utterances than in V+ utterances. Furthermore, it appears that voicing-related (and stress-related) differences in jaw steady-state duration are greater in utterances which do not have other factors operating to reduce their duration. Thus, for EB and BW, the change in steady-state duration across voicing conditions was greater in S+ utterances than in S- utterances and greater in utterances containing final fricatives than final stops. Similar interactions were reported in the analysis of total vowel duration.

Influence of Stress and Voicing on Jaw Raising

Jaw Raising Slope

Stress had a clear and consistent influence on the slope of the jaw raising gesture; S+ utterances demonstrated significantly steeper slopes than S- utterances for each speaker (ps < .0001). A significant stress x voicing interaction mediated the stress main effect for EB (p= .0004). For this speaker, the difference in slope across stress conditions was greater in V- utterances than in V+ utterances.

Voicing also significantly influenced the slope of the jaw raising gesture; V- utterances demonstrated steeper raising slopes than V+ utterances for each speaker (p< .0001 (VS)(EB);p= .0004 (BW)). A significant voicing x manner interaction mediated the voicing main effect for VS. For this speaker, the voicing-related difference in raising slope was greater in utterances containing final stops than final fricatives. As already mentioned, a significant stress x voicing interaction mediated the voicing main effect for EB. The change in slope across voicing conditions was greater in S+ utterances than in S- utterances for this speaker.

Jaw Raising Duration

Stress did not have a consistent influence on the duration of the jaw raising gesture. For VS, S+ utterances displayed significantly longer raising durations than S- utterances (p< .0001). S+ utterances displayed longer raising durations than S- utterances for BW also, although the stress main effect was not significant for this speaker (p= .0189). For EB, the stress main effect was significant but in the opposite direction to the pattern seen for VS and BW. S+ utterances had significantly shorter raising durations than S- utterances for EB (p= .0004). A significant stress x voicing interaction mediated the stress main effect for VS (p= .0013). Stress-related differences in raising duration were greater in V+ utterances than V- utterances for this speaker.

One of the clearest and most consistent effects of voicing on jaw articulation was on the duration of the raising gesture. For all speakers, V+ utterances demonstrated significantly greater raising durations than V-utterances ($ps < .0001$). Significant voicing x manner interactions were present in the analyses for BW and EB ($p < .0001$ (BW); $p = .0002$ (EB)). For these two speakers, the difference in raising duration across voicing conditions was greater in utterances containing final fricatives than final stops. As already mentioned, a significant stress x voicing interaction was present for VS ($p = .0013$). Voicing had a greater effect on raising duration for S+ utterances than S- utterances for this speaker. Once again, these interactions are similar in nature to the interactions reported in the analysis of total vowel duration, i.e., stress and voicing effects on duration are greatest in utterances where other factors are not operating to reduce duration.

Jaw Position at Vowel Offset

Stress influenced jaw position at vowel offset; S+ utterances displayed higher offset positions than S- utterances. The stress main effect was consistent in direction across speakers but statistically significant for speakers VS and BW only ($p < .0001$ (VS,BW); $p = .0824$ (EB)). For VS, stress x voicing and stress x vowel interactions were also significant ($p = .0008$ and $p = .0017$). The stress effect, while consistent in direction across voicing and vowel conditions, was greater for V+ utterances than V- utterances and greater for /æ/ utterances than /a/ utterances for this speaker. For BW, a significant stress x manner interaction was present. For this speaker, S+ utterances showed higher offset positions than S- utterances within each manner category with a larger stress-related difference for utterances containing final fricatives than final stops.

Voicing did not affect jaw position at vowel offset in a consistent manner across speakers. The main effect of voicing on offset position was significant for BW only ($p < .0001$). V+ utterances displayed higher positions at vowel offset than V- utterances for this speaker. No significant voicing main effects were observed in the analyses of jaw offset position for VS and EB. However, significant voicing x manner interactions were present for each of these speakers ($p < .0001$ (VS); $p = .0007$ (EB). For these two speakers, V+ utterances demonstrated higher offset positions than V- utterances for utterances containing final fricatives while V- utterances had higher offset positions for utterances containing final stops. A significant voicing x stress interaction for VS ($p = .0008$) reflected the fact that voicing effects were in opposite directions across stress conditions for this speaker. Within S+ utterances, V+ utterances demonstrated higher offset positions than V- utterances. For S- utterances, V- utterances had higher offset positions than V+ utterances.

Stress effects on jaw movement can be summarized as follows. S+ utterances were produced with more rapid jaw lowering gestures that were of greater duration than seen for S- utterances. These increases in velocity and duration allowed for more extreme (lower) articulatory positions at lowering offset for S+ utterances. Steady-state positions (following lowering offset and preceding raising onset) were maintained for greater durations in S+ utterances than in S- utterances. Finally, S+ utterances demonstrated more rapid jaw raising gestures allowing higher jaw positions to be attained at vowel offset. The durations of jaw raising gestures were not influenced by stress in a consistent manner across the three speakers.

Turning to final-consonant voicing effects, jaw lowering gestures were slightly more rapid (i.e., had steeper slopes) and reached more extreme (lower) positions at lowering offset for V- utterances than V+ utterances. The voicing-related differences in jaw lowering slope and position at lowering offset were not as large as the stress-related differences in these same variables. Although jaw steady-state positions were more extreme (lower) in V- utterances, steady-state positions were maintained for greater durations in V+ utterances. Jaw raising gestures were more rapid in V- utterances but durationally longer in V+ utterances. Voicing consistently influenced jaw position at vowel offset for only one speaker (BW) with V+ utterances demonstrating higher positions at vowel offset than V- utterances.

Although final-consonant voicing did have some influence on articulation in the early (jaw-lowering) portions of the utterances, these effects were not as large as stress effects on these early portions. Voicing effects on articulation became larger and more consistent in later portions of the utterances. Stress, on the other hand, appears to have influenced jaw articulation in a less time-dependent manner, affecting early and later portions of the utterances approximately equally. The voicing results contrast with the stress results in another important way. Recall that in examining stress the utterances with a greater total vowel duration (i.e., the S+ utterances) demonstrated more rapid lowering and lower positions at lowering offset. In the voicing results, exactly the opposite pattern was observed. Here the temporally shorter V- utterances display steeper slopes and lower positions at lowering offset. This contrast between stress and voicing effects on production is also evident in the formant data described in the next section.

Results III

Influence of Stress and Final-Consonant Voicing on Formant Structure

This section examines the effects of stress and final-consonant voicing on F1 structure and on F2 peak frequency during vowel production. F1 structure was examined in terms of eight descriptive properties. F1 initial transitions were examined for onset frequency, transition slope, transition duration, and offset frequency (steady-state frequency as defined in the Method section). The F1 steady-state region was examined for frequency and duration. F1 final transitions were examined for final transition onset frequency (equivalent to steady-state frequency), transition slope, transition duration, and offset frequency.

Mean values for S+ versus S- utterances for each of these variables are listed for each speaker in the lefthand portion of Table II. Mean values for V+ utterances versus V- utterances appear in the righthand portion of the table. The F1 mean frequencies and mean durations for S+ and S- utterances given Table II are displayed graphically in the panels of Figure 7. Each panel of the figure contains two lines representing mean F1 values for stressed versus unstressed utterances for a particular speaker. The initial rising transitions in each panel are based on mean F1 onset frequencies and mean initial transition durations. The steady-state regions represent mean steady-state frequencies and steady-state durations. The final transitions represent mean final transition durations and offset frequencies. The mean frequencies and mean durations for V+ and V- utterances listed in Table II are represented in the panels of Figure 8 using the same format as used in Figure 7. The mean slopes of the initial and final transitions listed in Table II were not used in constructing Figures 7 and 8. However, the slopes of the transitions in these figures are representative of these mean slopes in most instances.

Insert Table II, Figure 7, and Figure 8 about here

One goal of the present investigation was to determine the extent to which stress-related and voicing-related changes in the articulatory data are related to corresponding changes in the formant data. For example, we expected that more rapid jaw movement would produce steeper formant transitions, more extreme jaw lowering would produce more extreme formant frequencies, etc. Having already presented the articulatory results for stress and voicing effects on jaw movement, the present section will refer to these results and describe the extent to which stress and voicing effects on formant structure do or do not conform to expectations based on the articulatory data.

Influence of Stress and Voicing on F1 Initial Transitions

F1 Onset Frequency

In our earlier analyses, S+ utterances demonstrated slightly higher jaw positions at voicing onset than S- utterances for each speaker. The small differences in jaw position across stress levels were nonsignificant in each case. These articulatory results led to the prediction that F1 onset frequency would show little change across stress levels. Based on the expectation that lower jaw positions correspond to higher F1 frequencies, a tendency for S+ utterances to show slightly lower onset frequencies would be anticipated. The expectation of no large stress effects on F1 onset frequency was supported. The main effect of stress was significant for BW only, with S+ utterances showing higher onset frequencies than S- utterances (p= .0034). The mean values for VS and EB were also in the direction of S+ utterances showing higher onset frequencies. Thus, the expected relationship between jaw position on F1 frequency was not entirely supported in this instance.

With regard to voicing, the articulatory data demonstrated little voicing effect on jaw position at voicing onset. Thus F1 frequencies at voicing onset were not expected to show consistent differences across voicing conditions. However, voicing did have a consistent influence on F1 onset frequency with V+ utterances showing significantly lower onset frequencies than V- utterances for each speaker (p< .0001 (VS,EB);p= .0034 (BW)). The results suggest that jaw position at voicing onset may not be highly correlated with F1 onset frequency.

F1 Initial Transition Slope

The articulatory data demonstrate more rapid initial jaw lowering (steeper lowering slopes) for S+ utterances than for S- utterances for all speakers. F1 initial transition slopes were expected to be steeper in S+ utterances than S- utterances as a result. F1 initial transition slopes were significantly steeper for S+ utterances for BW and EB (p= .0033 (BW);p< .0001 (EB). The stress main effect was in the same direction but nonsignificant for VS (p= .2484). A significant vowel x stress interaction was present for VS (p= .0021). For this speaker, S+ utterances containing /a/ demonstrated steeper slopes the S- /a/ utterances while S- utterances containing /æ/ demonstrated slightly steeper slopes than S+ /æ/ utterances. Thus the expectation of S+ utterances demonstrating steeper F1 initial transitions was

Table II. Least squares estimated means on F1 and F2 variables for S+ versus S- utterances and V+ versus V- utterances.

formant property	Speaker	stressed utterances	unstressed utterances	stressed - unstressed		voiced utterances	voiceless utterances	voiced - voiceless	
F1 frequency at voicing onset (Hz)	VS	621	609	12		605	625	-20	***
	BW	598	587	11	*	587	598	-11	*
	EB	568	560	8		553	575	-22	***
F1 initial transition slope (Hz/ms)	VS	3.03	2.81	0.22		2.69	3.15	-0.46	
	BW	2.13	1.65	0.48	*	1.98	1.81	0.17	
	EB	3.66	2.51	1.15	***	3.25	2.92	0.33	
F1 initial transition duration (ms)	VS	86	72	14	***	83	74	9	
	BW	86	65	21	***	68	83	-15	**
	EB	49	47	2		43	53	-10	
F1 steady-state frequency (Hz)	VS	810	787	23	***	780	816	-36	***
	BW	766	667	99	***	694	739	-45	***
	EB	713	658	55	***	665	705	-40	***
F1 steady-state duration (ms)	VS	105	76	29	***	100	80	20	***
	BW	72	68	4		82	57	25	***
	EB	93	85	8		96	82	14	
F1 final transition slope (Hz/ms)	VS	-3.66	-3.95	0.29		-4.20	-3.41	-0.79	
	BW	-3.97	-2.32	-1.65	***	-3.48	-2.81	-0.67	
	EB	-2.06	-1.79	-0.27		-2.08	-1.78	-0.30	
F1 final transition duration (ms)	VS	30	28	2		36	22	14	***
	BW	35	35	0		50	19	31	***
	EB	38	35	3		57	16	41	***
F1 frequency at vowel offset (Hz)	VS	699	662	37	***	633	728	-95	***
	BW	629	579	50	***	536	672	-136	***
	EB	646	595	51	***	567	675	-108	***
F2 peak frequency (Hz)	VS	1552	1524	28	***	1550	1526	24	***
	BW	1409	1335	74	***	1360	1384	-24	*
	EB	1484	1414	70	***	1454	1444	10	

*	$p < .004$
**	$p < .001$
***	$p < .0001$

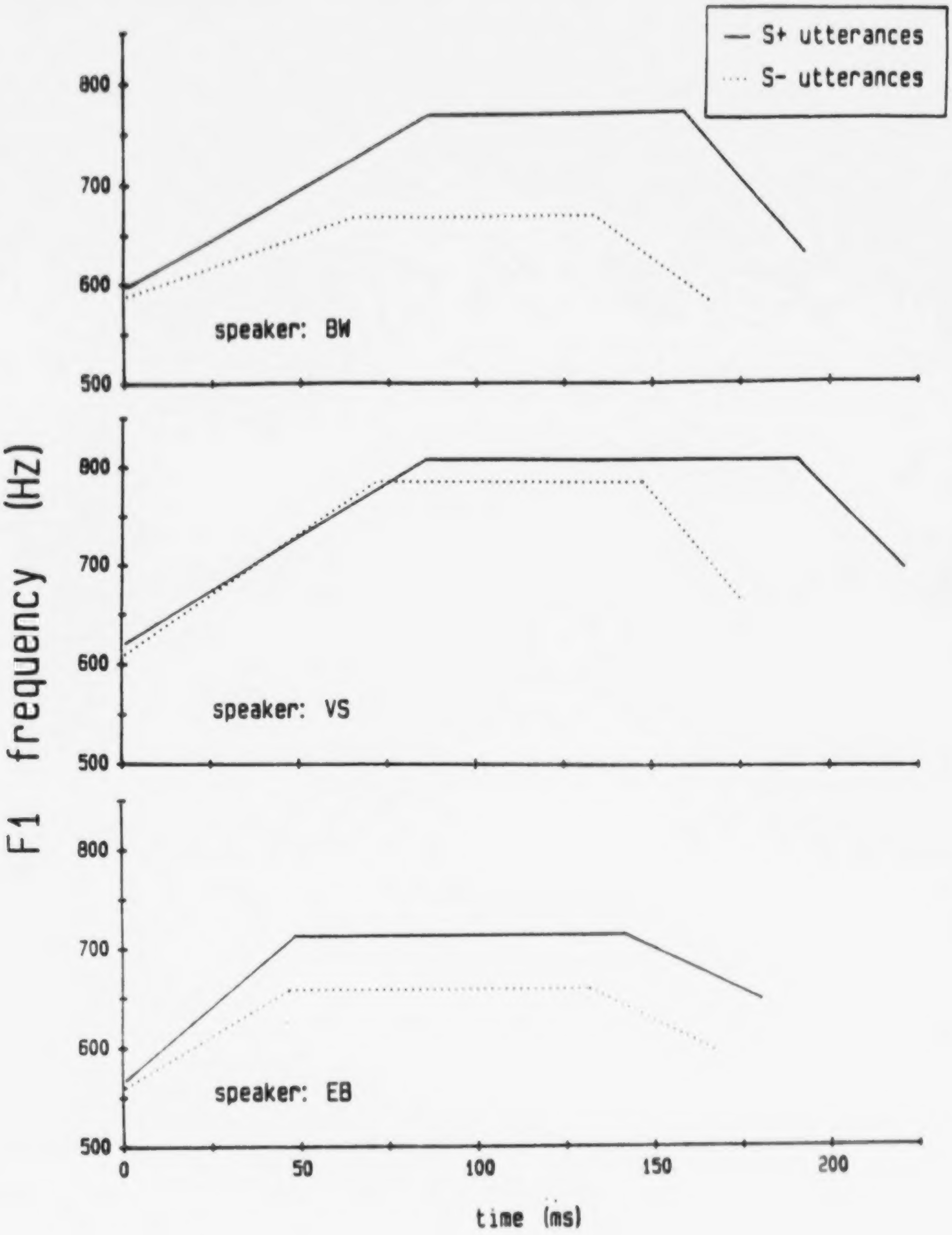


Figure 7. F1 frequency plots for stressed (S+) versus unstressed (S-) utterances. Based on mean frequencies and mean durations listed in Table II.

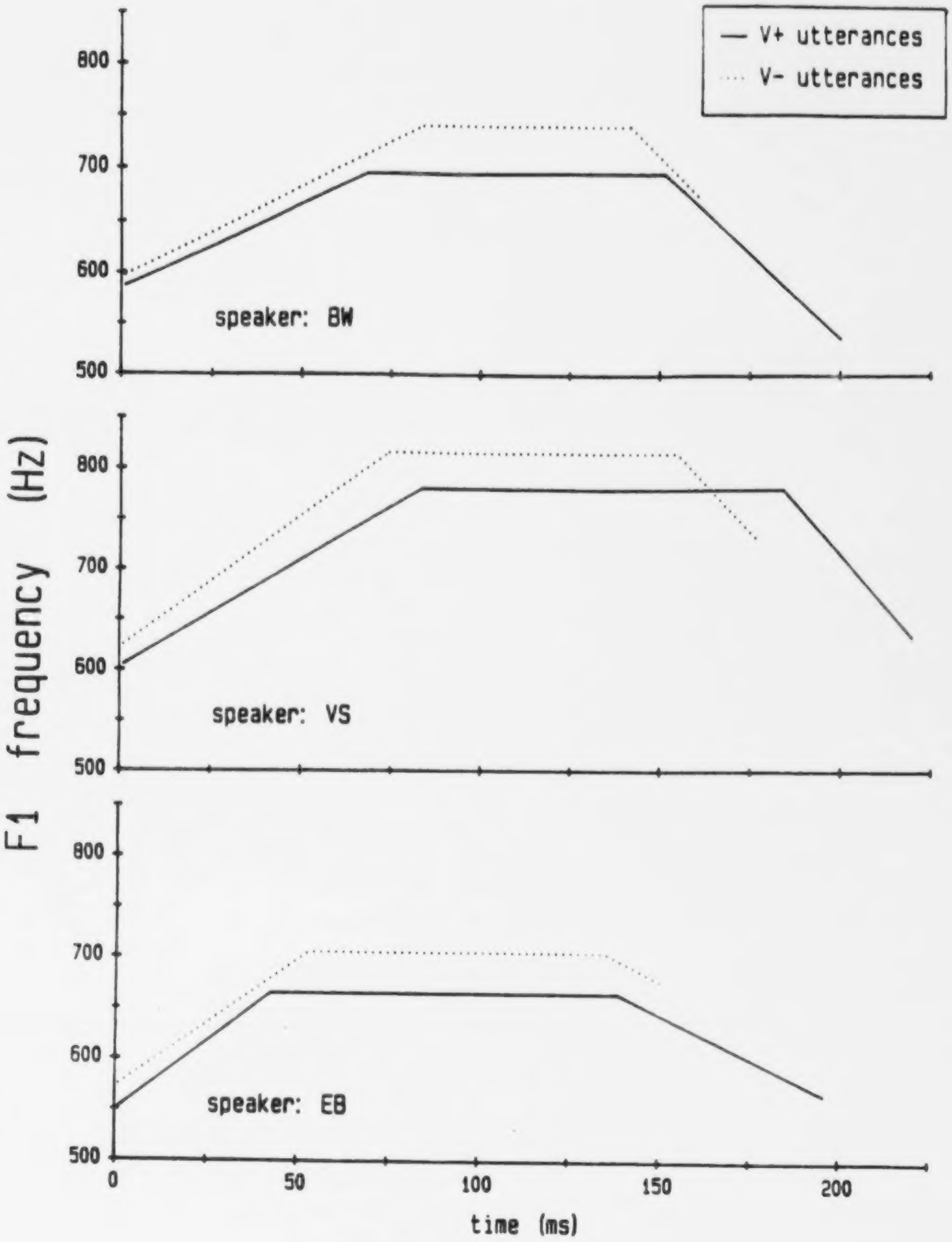


Figure 8. F1 frequency plots for utterances containing voiced (V+) versus voiceless (V-) final consonants. Based on mean frequencies and mean durations listed in Table II.

borne out for two of the three speakers and for /a/ utterances from the third speaker.

Voicing did not have as large or as consistent an influence on jaw lowering slopes as stress. However, V- utterances did tend to demonstrate more rapid jaw lowering (steeper lowering slopes) than V+ utterances (for EB in particular). It was therefore expected that V- utterances would demonstrate steeper F1 initial transitions than V- utterances. The formant data did not support this expectation. Voicing did not have a significant influence on F1 initial transition slopes for any of the speakers. Thus, while the large and consistent differences in jaw lowering slopes observed across stress conditions appeared to correlate with differences in F1 initial transition slopes, the smaller changes in jaw lowering slopes seen across voicing conditions were not reflected in the formant data.

F1 Initial Transition Duration

Jaw lowering durations were greater in S+ utterances than S- utterances. This led to the prediction that F1 initial transitions would also be longer in S+ utterances. Initial transitions were significantly longer for S+ utterances than S- utterances for VS and BW ($p < .0001$ (VS,BW)). S+ utterances showed only slightly longer initial transition durations than S- utterances for EB. The stress main effect did not approach significance for this speaker.

Voicing did not significantly influence jaw lowering durations for any of the speakers and was therefore not expected to influence F1 initial transition durations. A significant voicing effect on F1 initial transition duration was present for BW ($p = .0003$) with V- utterances demonstrating longer durations than V+ utterances. This main effect was mediated by a significant voicing x vowel interaction ($p < .0001$). For BW, V+ utterances containing /a/ demonstrated longer F1 initial transitions than V- /a/ utterances while V- utterances containing /æ/ were longer than V+ /æ/ utterances. Thus voicing did not have a consistent main effect on F1 initial transition duration for any of the three speakers.

F1 Initial Transition Offset Frequency (steady-state frequency)

Jaw positions at lowering offset (jaw steady-state positions) were lower in S+ utterances than S- utterances for all speakers. These lower jaw positions were expected to correlate with higher F1 steady-state frequencies. This expectation was supported for all three speakers with S+ utterances demonstrating higher F1 steady-state frequencies than S- utterances ($p < .0001$ for each speaker). A significant stress x voicing interaction was also observed for BW ($p < .0001$). Stress had a greater influence on steady-state frequency for V- utterances than V+ utterances for this speaker.

Turning to voicing, jaw steady-state positions were lower for V- utterances than V+ utterances for each speaker. V- utterances were therefore expected to demonstrate higher F1 steady-state frequencies than V+ utterances. This expectation was also borne out for all three speakers. For each speaker, V- utterances demonstrated significantly higher F1 steady-state frequencies than V+ utterances ($p < .0001$ for each speaker). A significant stress x voicing interaction was present in the analysis of F1 steady-state frequency for BW. For this speaker, the influence of voicing on F1 steady-state frequency was greater for S+ utterances than S- utterances. A significant voicing x vowel interaction was present in the analysis of F1 steady-state frequency for VS ($p = .0013$). For this speaker, the voicing effect was greater for utterances containing /æ/ than /a/.

To summarize, differences in F1 initial transition structure across stress levels were fairly well correlated with stress effects on jaw lowering. The more rapid, durationally longer jaw lowering gestures demonstrated by S+ utterances apparently contributed to the more rapid, longer F1 initial transitions seen for these same utterances. The lower jaw positions demonstrated by S+ utterances appear to be related to the higher F1 steady-state frequencies which S+ utterances also demonstrated. Voicing-related changes in F1 initial transition structure were not well predicted from the articulatory data in most cases. Lower F1 onset frequencies demonstrated by V+ utterances did not correspond to significant voicing-related differences in jaw position at voicing onset. No consistent voicing-related differences in F1 initial transitions slopes or durations were seen, although the articulatory data suggested that transitions slopes would be slightly more rapid for V- utterances. The one case in which voicing-related changes in articulation showed a clear relationship to changes in formant structure concerned jaw position at lowering offset (steady-state position) and F1 frequency at initial transition offset (steady-state frequency). V- utterances displayed lower jaw steady-state positions and higher F1 steady-state frequencies than V+ utterances for all three speakers.

Influence of Stress and Voicing on F1 Steady-state Region

F1 Steady-state Frequency and Duration

As already described, F1 steady-state frequencies were higher in S+ utterances than in unstressed utterances. S+ utterances were also expected to demonstrate longer F1 steady-state durations based on the articulatory data where the period between jaw lowering offset and raising onset (jaw steady-state duration) was greater for S+ utterances. Longer F1 steady-state durations for S+ utterances were clearly present for VS ($p < .0001$). While S+ utterances displayed longer durations than S- utterances for BW and EB also, the stress main effect was not significant for either speaker ($p = .4531$ (BW); $p = .1068$ (EB)). Thus the pattern of S+ utterances showing longer steady-state durations than S- utterances was more reliable in the articulatory data than in the formant data.

While V- utterances demonstrated higher F1 steady-state frequencies than V+ utterances, the articulatory data suggested that F1 steady-state durations would be shorter for these utterances than for V+ utterances. That is, the longer jaw steady-state durations demonstrated by V+ utterances were expected to be correlated with longer F1 steady-state durations. V+ utterances demonstrated significantly longer F1 steady-state durations than V- utterances for VS and BW ($p < .0001$ (VS,BW)). The main effect of voicing was nonsignificant for EB ($p = .0079$) although V+ utterances again showed longer durations than V- utterances. Several interactions involving voicing complicated these results. Significant voicing x vowel interactions were present for both BW and EB ($p < .0001$(BW);$p = .0003$ (EB)). For each speaker, V+ utterances containing /æ/ were clearly longer in duration than corresponding V- utterances while V+ utterances were slightly shorter than V- utterances for utterances containing /a/. In addition, a significant voicing x manner interaction was seen for EB ($p = .0002$). For EB, F1 steady-state durations were clearly longer in V+ utterances containing final stops than in corresponding V- utterances. However, V- utterances containing fricatives had slightly longer F1 steady-state durations than V+ utterances containing fricatives. Thus while V+ utterances tended to demonstrate longer steady-state durations than V- utterances, the pattern was not completely consistent. The pattern of V+ utterances showing longer steady-state regions, while present in both the articulatory and formant data, was more consistent and reliable in the jaw movement data than in the F1 data.

Influence of Stress and Voicing on F1 Final Transitions

F1 Final Transition Slope

The slopes of jaw raising gestures were steeper in S+ utterances than S- utterances for each speaker, suggesting more rapid raising gestures for S+ utterances. These more rapid raising gestures were expected to be correlated with steeper F1 final transitions for S+ utterances. This prediction was supprted for BW; S+ utterances demonstrated significantly steeper F1 final transitions than S- utterances (p< .0001). Stress did not significantly influence F1 final transition slope for VS or EB. A significant stress x vowel interaction for BW (p= .0007) was due to a larger stress effect for /a/ utterances than / / utterances. A significant stress x voicing interaction was present in the analysis of final transition slope for EB (p= .0008). For this speaker, the direction of the stress effect varied across voicing categories. S+ utterances showed steeper final transitions for V- utterances while S- utterances showed steeper transistions for V+ utterances. The expectation of steeper F1 final transitions being correlated with S+ utterances was not consistently supported across speakers.

V- utterances demonstrated significantly steeper jaw raising slopes than V+ utterances for all speakers. Therefore, it was expected that V- utterances would demonstrate steeper F1 final transitions than V+ utterances. This expectation was not supported; voicing did not have a significant effect on F1 final transition slopes for any of the speakers. A significant voicing x manner interaction was observed in the analysis of F1 final transition slope for BW (p= .0022). For BW, V+ utterances containing final fricatives showed steeper slopes than corresponding V- utterances while V- utterances containing final stops showed steeper slopes than corresponding V+ utterances. A significant stress x voicing interaction was seen in the analysis of F1 final transition slope for EB. For this speaker, V+ utterances showed steeper final transitions than V- utterances in the S- context while V- showed steeper slopes in the S+ condition. The expectation of steeper F1 final transition slopes being associated with V- utterances was not supported.

F1 Final Transition Duration

Stress effects on jaw raising duration were extremely variable across speakers (see Table I) which led to an expectation of inter-speaker variability in terms of stress effects on F1 final transition durations. Little variability was seen however. Stress did not significantly influence F1 final transition duration for any speaker. For each speaker, final transition mean durations were slightly, but nonsignificantly, longer for S+ utterances than S- utterances.

Voicing had consistent effects on jaw raising durations with V+ utterances showing significantly longer raising duration than V- utterances for each speaker. These longer raising durations were expected to be correlated with longer F1 final transitions for V+ utterances. This expectation was supported across speakers. V+ utterances demonstrated significantly longer F1 final transition durations than V- utterances for each speaker (p< .0001 for each speaker). Significant voicing x manner and voicing x vowel interactions were present in the analyses of F1 final transition duration for BW and EB (p< .0001 for the voicing x manner interactions and the voicing x vowel interaction for EB; *p*= .0003 for the voicing x vowel interaction for BW). For each speaker, voicing had a greater effect on duration for utterances containing /a/ than /æ/, and a greater effect for utterances containing final fricatives than final stops.

F1 Frequency at Vowel Offset

Jaw positions at vowel offset were higher for S+ utterances than S- utterances for all speakers (significantly higher for VS and BW). The expectation of low jaw positions correlating with high F1 frequencies would therefore suggest that F1 final transition offset frequencies would be lower in S+ utterances than S- utterances. However, exactly the opposite pattern was observed. Final transition offset frequencies were significantly higher for S+ utterances than S- utterances for each speaker (*p*< .0001 for each speaker). A significant stress x vowel interaction was also present for VS (*p*= .0033), reflecting a larger stress effect for /a/ utterances than /æ/ utterances.

Voicing had a significant main effect on jaw position at vowel offset for BW only. V+ utterances displayed higher offset positions than V- utterances for this speaker. It was therefore expected that F1 offset frequencies might be lower for V+ utterances than V- utterances for BW and that this pattern might not be seen for the other speakers. However, V+ utterances showed lower F1 offset frequencies than V- utterances for all speakers (*p*< .0001 for each speaker). Significant voicing x manner interactions occurred in the analyses of F1 offset frequency for BW and EB (*p*s < .0001). For these two speakers, voicing had a larger effect on offset frequency for utterances containing final stops than final fricatives. The stress and voicing results suggest that jaw position and F1 frequency, while consistently related to each other during the central, steady-state portions of the vowels, are not closely related at vowel onset or vowel offset.

F2 Peak Frequency

It has already been pointed out that F2 is traditionally thought of as being related to the front-back position of the articulators (particularly the tongue). Since the present study did not include articulatory measurements in the horizontal plane, the F2 results were not compared with the articulatory data.

F2 peak frequency means for S+ versus S- utterances and for V+ versus V- utterances are listed in Table II. F2 peak frequencies were significantly higher in S+ utterances than in S- utterances for each speaker (*p*s < .0001). A significant stress x vowel interaction mediated the stress main effect for VS (*p*< .0001). For this speaker, S+ utterances containing /a/ showed slightly lower F2 frequencies than S- utterances containing /a/ (means differed by less than 5 Hz across stress conditions). Thus the pattern of S+ utterances showing higher F2 peak frequencies than S- utterances, while consistent across vowels for speakers BW and EB, was only true for utterances containing /æ/ for speaker VS.

Final-consonant voicing did not have a consistent main effect on F2 peak frequencies. For VS, V+ utterances showed significantly higher frequencies than V- utterances (*p*< .0001). For BW, the voicing main effect was also significant but in the opposite direction; V- utterances showed higher F2 frequencies than V+ utterances (*p*= .0004). The voicing main effect was not signficant for speaker EB. However, a fairly consistent voicing-related pattern was seen within each of these analyses. Significant voicing x vowel interactions were present for each speaker ($p< .0001$ (BW, VS);$p= .0011$ (EB)). For each speaker, V+ utterances containing /æ/ demonstrated higher F2 frequencies than corresponding V- utterances. For VS, voicing had a neglible effect on F2 frequency for utterances containing /a/; the difference in means across voicing conditions was less than 2 Hz. For BW and EB, the voicing effect was in the opposite direction for utterances containing /a/ compared to /æ/ utterances. For these two speakers, V- utterances containing /a/ demonstrated higher F2 peak frequencies than corresponding V+ utterances.

Summarizing stress effects on formant structure, we found that S+ utterances displayed steeper, durationally longer F1 initial transitions than S- utterances. These more rapid, lengthier initial transitions produced higher F1 steady-state frequencies in S+ utterances. Stress effects on F1 final transition characteristics were less consistent, although S+ utterances demonstrated significantly higher F1 offset frequencies than S- utterances for each speaker. F2 peak frequencies were higher in S+ utterances than in S- utterances for BW and EB, and for VS' utterances containing /æ/.

Summarizing the voicing effects, we found that V- utterances displayed higher F1 onset frequencies and higher steady-state frequencies than V+ utterances. F1 steady-state durations and final transition durations were longer in V+ utterances than unstressed utterances. F1 offset frequencies were much lower for V+ utterances than V- utterances. Contrasting with the stress results, decreases in vowel duration due to final-consonant voicing did not lead to neutralization of F1 frequencies toward more central values. In fact, the shorter, V- utterances actually demonstrated more extreme (higher) F1 frequencies than the longer, V+ utterances. This contrast between stress and voicing effects was also noted in the articulatory data with respect to jaw steady-state position. Voicing interacted with vowel in influencing F2 peak frequency. For utterances containing /æ/, V+ utterances showed higher F2 frequencies than V- utterances for all speakers. For utterances containing /a/, V- utterances showed higher F2 frequencies than V+ utterances for two of the three speakers.

General Discussion

Previous research has shown that vowels in unstressed syllables are shorter in duration and often demonstrate more centralized (/ʌ/-like) steady-state formant frequencies than stressed vowels. This change in steady-state frequencies is referred to as vowel reduction and has been reported in a number of acoustic studies (Delattre, 1969; Gay, 1978; Harris, 1978; Lindblom, 1963; Tiffany, 1959). The present data replicate these earlier findings concerning stress-related vowel reduction. Furthermore, the present results provide information on how this stress-related vowel reduction was achieved in terms of the overall restructuring of the articulatory and acoustic patterns.

In terms of articulation, stressed utterances demonstrated longer, more rapid jaw lowering gestures than unstressed utterances. The greater duration and velocity of these lowering gestures resulted in lower steady-state jaw positions for stressed utterances. These more extreme steady-state positions were maintained for longer durations than in unstressed utterances. Finally, raising gestures were more rapid and reached higher positions at vowel offset in stressed utterances than in unstressed utterances. Stressed gestures were of greater magnitude for most of the articulatory dimensions examined. In general, the presence of stress increased the duration, velocity, and spatial extensiveness of jaw movement. Several of the stress effects on jaw movement reported here have been reported previously in the literature. Stress-related increases in articulator lowering velocity have previously been reported for both the jaw (Stone, 1981) and lower lip (Kelso, Vatikiotis- Bateson, Saltzman, & Kay, 1985). Kelso et al. (1985) also reported greater articulatory lowering durations in stressed utterances which again agrees with the results reported here. The finding of more extreme articulatory positions being associated with S+ utterances has been reported in several studies (Kelso et al., 1985; Kent & Netsell, 1971; Kozhevnikov & Chistovich, 1965; Stone, 1981). Finally, Kent & Netsell (1971) and Kelso et al., (1985) also reported stress-related increases in articulator raising velocities which are in agreement with the present results.

The formant data were slightly less consistent in terms of stress effects. Nevertheless, the overall picture is once again one of longer, more rapid, and more extreme formant movement in the stressed context. Specifically, F1 initial transitions tended to be steeper and of greater duration in S+ utterances, producing higher F1 steady-state frequencies. F1 steady-state durations were also of greater duration in S+ utterances compared to S- utterances. Stress effects on F1 final transition characteristics were not as consistent as stress effects on jaw raising characteristics. The only consistent stress effect on F1 final transitions concerned offset frequency; S+ utterances demonstrated consistently higher offset frequencies than S- utterances.

The articulatory data demonstrate that jaw movement is less rapid and both spatially and temporally less extensive in S- utterances. This reduction in the overall magnitude of the articulatory gestures is clearly related to the neutralization of F1 seen in the formant data. The articulatory data were not expected to relate directly to the F2 peak frequency data. However, it was anticipated that reduction in F2 frequencies towards more central (/ʌ/-like) values would accompany stress removal. This expectation was only partially supported by our analyses. For each speaker, S+ utterances containing /æ/ demonstrated higher F2 peak frequencies than corresponding S- utterances. The results for /æ/ were consistent with the expectation that vowel reduction accompanies stress removal. F2 is clearly higher for /æ/ than /ʌ/ in the Peterson and Barney (1952) data. Vowel reduction should therefore result in a decrease in F2 frequency for /æ/, as was found in the present study. However, the Peterson and Barney (1952) data show /a/ as having a fairly central F2 steady-state frequency only slightly lower than that of /ʌ/. Vowel reduction would therefore be expected to produce very little change in F2 frequency for /a/. This result was observed for speaker VS only.

In his well-known study on'vowel reduction, Lindblom (1963) suggested that the relationship between vowel duration and vowel reduction described above is universal. That is, a decrease in vowel duration (due to the destressing of an utterance, an increase in speaking rate, etc.) was assumed to lead directly to a reduction in vowel formant frequencies towards /ʌ/. In the present study, this relationship was not observed when final-consonant voicing is the factor influencing vowel duration (at least in terms of the F1 data). In addition, reports of vowel reduction accompanying voicing-related decreases in vowel duration could not be located in the acoustic literature (Lindblom did not examine utterances differing in final voicing in his 1963 paper). The presence of vowel reduction across stress levels and the absence of reduction across voicing conditions obviously requires a different sort of restructuring of the formant patterns in the two cases. These differences are also clearly reflected in the articulatory data.

V- utterances demonstrated slightly greater jaw-lowering slopes than V+ utterances, suggesting more rapid jaw lowering for V- utterances. This more rapid jaw lowering apparently contributed to the more extreme (lower) jaw steady-state positions demonstrated by V- utterances. More extreme jaw lowering for V- utterances has also been reported by Fujimura and Miller (1979). Steady-state jaw positions, while spatially more extreme for V- utterances, were maintained for greater durations in V+ utterances. Jaw-raising gestures were steeper in slope but briefer in duration in V- utterances compared to V+ utterances. The pattern of V- utterances demonstrating more rapid articulatory raising than V+ utterances has been reported previously for both the jaw (Fujimura & Miller, 1979) and lower lip (Chen, 1970). The influence of voicing on articulatory properties was small during early portions of the utterances (i.e., during jaw-lowering) and increased during later-occuring portions.

Although final-consonant voicing did affect F1 frequency at voicing onset (V+ utterances showed lower onset frequencies), voicing effects on F1 structure were more prevalent in later-occuring portions of the test utterances. F1 steady-state regions demonstrated higher frequencies and briefer durations in V- utterances compared to V+ utterances (see Wolf (1978) for similar findings with respect to F1 steady-state frequency). F1 final transitions were briefer in duration and terminated at much higher frequencies for V- utterances.

F1 steady-state frequencies were not neutralized towards /ʌ/ in the shorter, V- utterances. However, changes in F2 peak frequencies across voicing conditions were consistent with a vowel neutralization explanation. As already described, vowel neutralization would require a lowering of F2 for utterances containing /æ/ and a slight rise in F2 for utterances containing /a/. While final-consonant voicing had almost no effect on F2 frequency for VS' utterances containing /a/, more neutral F2 frequencies were seen for the V- utterances compared to V+ utterances in all other instances.

V- utterances were considerably shorter than V+ utterances in total vowel duration. However, initial jaw lowering durations and F1 initial transition durations gave no evidence of being briefer in V- utterances. In order for jaw-lowering durations and F1 initial transition durations to be similar across voicing conditions, V- utterances necessarily contributed a greater proportion of total vowel duration to jaw lowering and to initial F1 transitions. This result is consistent with earlier acoustic and perceptual data suggesting that voiceless utterances contain proportionally longer initial transitions than voiced utterances (Fitch, 1981; Soli, 1982). These results are again in agreement with the general observation that final-consonant voicing has greater effects on the final portions of vowels than on the initial portions while stress has a more global influence on articulatory movements and formant structure throughout vowel production. Stress is generally described as a suprasegmental feature of speech which is overlaid on the segmental articulatory pattern. Final-consonant voicing, on the other hand, is described as a segmental feature of a given consonant. Thus the greater influence of stress on articulatory movements during the early portions of the vowels and the increasing influence of voicing on production during later portions of the vowels is not surprising.

Parker (1974) has suggested that vowel termination characteristics contain final-consonant voicing information with gradual termination cuing a voiced consonant and abrupt termination cuing voicelessness. The longer, less rapid articulatory raising gestures demonstrated by V+ utterances in the present data fit nicely with this description. Other researchers have also focused on vowel termination characteristics as an important source of voicing information (Hillenbrand, Ingrisano, Smith, & Flege, 1984; Revoile, Pickett, Holden, & Talkin, 1982; Walsh, Parker, & Miller, 1985; Walsh & Parker, 1981; Wardrip-Fruin, 1982; Wolf, 1978). Several of these researchers have reported both acoustic and perceptual data suggesting F1 offset frequency as a voicing cue with lower offset frequencies cuing voicing (Hillenbrand et al., 1984; Revoile et al., 1982; Wolf, 1978). These results are also in agreement with the present data.

In the present data, decreases in vowel duration due to stress removal had very different effects on jaw movements and formant structure than decreases in duration due to devoicing of a following consonant. Decreases in vowel duration due to stress were accompanied by reductions in the speed, duration, and extensiveness of articulatory gestures. This "reduction" is also evident at the acoustic level with less rapid formant transitions and more central (/ʌ/-like) F1 steady-state frequencies occurring in unstressed utterances. Vowel duration differences due to final-consonant voicing did not reflect this pattern. That is, the durationally shorter voiceless utterances did not show any tendency towards reduction at either the articulatory or acoustic level (in terms of F1). In fact, in the present data, more rapid and more extreme articulatory lowering and more extreme F1 steady-state frequencies were observed for the durationally shorter, voiceless utterances than for the longer, voiced utterances.

Stress and final-consonant voicing influence F1 structure differently. Consequently, vowel duration and F1 structure, when jointly examined, may uniquely specify both the stress and final-consonant voicing information for a given utterance. First, extremely long vowel durations are associated with stressed utterances containing voiced final consonants since both factors are working to lengthen vowel duration. Likewise, extremely short durations are associated with unstressed utterances containing voiceless final consonants. However, vowel duration does not disambiguate the two remaining combinations of stress and voicing. S+ utterances containing voiceless final consonants and S- utterances containing voiced final consonants are fairly similar in total vowel duration. It is these two sets of utterances which differ the most in F1 structure. The stressed-voiceless utterances show the steepest F1 initial transitions and the most extreme F1 steady-state frequencies. In this case the presence of stress and the absence of final-consonant voicing work together to produce rapid transitions and high steady-state frequencies. The unstressed-voiced utterances have both factors working to reduce F1 initial transition slopes and to neutralize F1 steady-state frequencies. These utterances show the most gradual F1 transitions and most central steady-state frequencies. The large difference in F1 structure for these two types of utterances can be seen by comparing the F1 trajectories for the stressed /bap/ and unstressed /bab/ tokens in Figure 3. Taken together, vowel duration and F1 structure appear to unambiguously specify both stress and final-consonant voicing information.

A nearly identical pattern to the one just described can be observed in the articulatory data. As in the F1 data, the jaw movement data disambiguate the stressed-voiceless utterances from the unstressed-voiced utterances (the sets of utterances with similar vowel durations). Stressed-voiceless utterances demonstrate the greatest lowering velocities and lowest positions at lowering offset. In this case, stress and voicelessness are both working to increase lowering velocities and produce greater jaw movement. In unstressed-voiced utterances, both stress and voicing are working to reduce lowering velocities and overall jaw movement. These utterances demonstrate the smallest velocities and the highest positions at lowering offset. The difference in jaw movement across these two utterances types can be seen by comparing the the jaw trajectories for the stressed /bap/ and unstressed /bab/ tokens displayed in Figure 2.

The results of the present study suggest that vowel duration and vowel structure may both provide information for the perception of stress and final-consonant voicing. A detailed examination of vowel production at the articulatory and acoustic levels suggests that changes in vowel duration due to stress and final-consonant voicing are not accomplished by identical changes in articulatory and acoustic patterns. Rather it appears that the restructuring of the articulatory and acoustic patterns which accompany changes in vowel duration are specific to the factor producing the change in duration. The question of whether the stress-related and voicing-related differences in the articulatory and acoustic patterns reported here provide useful information for the perception of stress and voicing can best be addressed through perceptual experiments using synthetic stimuli which vary along dimensions suggested by the present data. These perceptual experiments are currently underway and reports of these findings should be forthcoming.

In summary, the present findings demonstrate that the simultaneous analysis of acoustic and articulatory data from an appropriate set of utterances can provide the beginnings of an articulatory explanation for acoustic variability due to stress and final-consonant voicing. Clearly, these techniques could be used to examine other linguistic and nonlinguistic factors known to influence articulatory and acoustic patterns. Factors such as intrinsic vowel duration, speaking rate, and position in utterance each have an important influence on vowel duration. An examination of the articulatory and acoustic correlates of vowel duration change associated with each of these factors could be carried out using techniques similar to those presented here. This research would be extremely valuable in determining whether each factor known to influence vowel duration has a consistent and unique influence on articulatory and acoustic structure.

References

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. Journal of the Acoustical Society of America, (51), 648-651.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. Phonetica, (22), 129-159.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. Journal of the Acoustical Society of America, (77), 2142-2156.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. International Review of Applied Linguistics, (7), 295-325.

Denes, P. (1955). Effect of duration on the perception of voicing. Journal of the Acoustical Society of America, (27), 761-764.

Derr, M. A., and Massaro, D. W. (1980). The contribution of vowel duration, FO contour, and frication duration as to the /juz/-/jus/ distinction. Perception and Psychophysics, (27), 51-59.

Fitch, H. L. (1981). Distinguishing temporal information for rate from temporal information for intervocalic stop consonant voicing. Status Report on Speech Research, (SR-65), 1-32.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. Journal of the Acoustical Society of America, (27), 155-158.

Fry, D. B. (1965). The dependence of stress judgments on vowel formant structure. In Proceedings of the 5th International Congress of Phonetic Sciences, Basal/New York: S. Karger, 1965.

Fujimura, O. and Miller, J. L. (1979). Mandible height and syllable-final tenseness. Phonetica, (36), 263- 272.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. Journal of the Acoustical Society of America, (63), 223-230.

Gulisano, V. (1982). PSP Physiological Signal Processing System. (Available from Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510)

Harris, K. S. (1978). Vowel duration and its underlying physiological mechanisms. Language and Speech, (21), 354-361.

Hillenbrand, J., Ingrisano, D. R., Smith, B. L., and Flege, J. E. (1984). Perception of the voiced-voiceless contrast in syllable-final stops. Journal of the Acoustical Society of America, (76), 18-27.

House, A. S., and Fairbanks, G. (1953). The influence of consonantal environment upon the secondary acoustical characteristics of vowels. Journal of the Acoustical Society of America, (25), 105-113.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (1985). A qualtitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, (*77*), 266-280.

Kent, R. D., and Netsell, R. (1971). Effects of stress contrasts on certain articulatory parameters. *Phonetica*, (*24*), 23-44.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, (*54*), 1102-1104.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics*, (*3*), 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, (*59*), 1208-1221.

Klatt, D. H., and Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception*. Berlin: Springer-Verlag.

Kozhevnikov, V. A., and Chistovich, L. A. (1965). *Speech: articulation and perception*. English translation distributed by Joint Publications Research Service, Washingtion, D.C.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, (*32*), 451-454.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, (*35*), 1773-1781.

Luce, P. A., and Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, (*78*), 1949-1957.

Mack, M. (1982). Voicing-dependent vowel duration in English and French: Monolingual and bilingual production. *Journal of the Acoustical Society of America*, (*71*), 173-178.

Morton, J., and Jassam, W. (1965). Acoustic correlates of stress. *Language and Speech*, (*8*), 159-181.

Nakatani, L. and Aston, C. H. (1978). Perceiving stress patterns of words in sentences. *Journal of the Acoustical Society of America*, (*63*), S55 (Abstract).

Oller, D. K. (1973). The effect of position in utterances on speech segment duration in English. *Journal of the Acoustical Society of America*, (*54*), 1235-1247.

Parker, F. (1974). The coarticulation of vowels and stop consonants. *Journal of Phonetics*, (*2*), 211-221.

Parmenter, C. E., and Trevino, S. N. (1936). Relative durations of stressed and unstressed vowels. *American Speech*, (*10*), 129-133.

Peterson, G. E., and Barney, H. L. (1952). Control methods in the study of vowels. *Journal of the Acoustical Society of America*, (*24*), 175-184.

Peterson, G. E., and Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, (*32*), 693-703.

Port, R. F. (1981) Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, (*69*), 262-274.

Port, R. F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics*, (*32*), 141-152.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. *Journal of the Acoustical Society of America*, (*51*), 1296-1303.

Revoile, S., Pickett, J. M., Holden, L. D., and Talkin, D. (1982). Acoustic cues to final stop consonant voicing for impaired- and normal-hearing listeners. *Journal of the Acoustical Society of America*, (*72*), 1145-1154.

Soli, S. (1982). Structure and duration of vowels together specify fricative voicing. *Journal of the Acoustical Society of America*, (*72*), 366-378.

Stone, M. (1981). Evidence for a rhythm pattern in speech production: Observations of jaw movement. *Journal of Phonetics*, (*9*), 109-120.

Summers, W. V. (1981). *Vowel duration and vowel structure in the cuing of lexical stress*. Unpublished master's thesis, University of Maryland, College Park, Maryland.

Tiffany, W. R. (1959). Nonrandom sources of variation in vowel quality. *Journal of Speech and Hearing Research*, (*2*), 305-317.

Walsh, T., Parker, F., and Miller, C. J. (1985). F1 transition as a cue to [+/- voice] in a following utterance-final stop. *Journal of the Acoustical Society of America*, (*77 Supp. 1*), S27-S28.

Walsh, T., and Parker, F. (1981). Vowel termination as a cue to voicing in post-vocalic stops. *Journal of Phonetics*, (*9*), 105-108.

Wardrip-Fruin, C. (1982). On the status of phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *Journal of the Acoustical Society of America*, (*71*), 187-195.

Westin, K., Buddenhagen, R. G, and Obrecht, D. H. (1966). An experimental analysis of the relative importance of pitch, quantity, and intensity as cues to phonemic distinctions in southen Swedish. *Language and Speech*, *9*, 114-126.

Wolf, C. G. (1978). Voicing cues in English final stops. *Journal of Phonetics*, (*6*), 299-309.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Preference Judgements Comparing Different Synthetic Voices*

John S. Logan and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This research was supported, in part, by NIH Research Grant NS-12179 and, in part by a contract with the Armstrong Aerospace Medical Research Laboratory (AFSC), Wright-Patterson AFB, OH. We thank Penny Mechley and Michael Dedina for their assistance in carrying out these experiments. The results of Experiment I were reported at the 111th meeting of the Acoustical Society of America, Cleveland, May 1986.

Abstract

Two experiments were carried out to study listener's preference judgements for synthetic speech from several different text-to-speech systems using an A/B paired comparison task. In both experiments, subjects heard a sentence produced by one system followed by the same sentence generated by another system. Subjects indicated which of the two voices they preferred and then furnished a confidence rating for their decision. In the first experiment, 40 Harvard Psychoacoustic sentences generated by the DECtalk, Prose 2000, and MITalk-79 systems were used as stimuli. In the second experiment, 92 Phoneme Specific sentences (Huggins & Nickerson, 1985) generated by the DECtalk, Prose 2000, and Infovox SA101 systems served as stimuli. The overall relationship among preference, response times, and confidence ratings was examined. In both experiments, a strong relation between preference and intelligibility was found. The remaining measures also appeared to be systematically related to judgements of subjective preference. The results are discussed in terms of the relationships among preference, naturalness, and intelligibility in the perceptual evaluation of natural speech and synthetic speech produced automatically by rule.

Preference Judgements Comparing Different Synthetic Voices

Most investigations of the quality of synthetic speech have concentrated primarily on segmental intelligibility since this aspect of perception is assumed to be central to its acceptance and use outside the laboratory. Despite the unquestionable role that segmental intelligibility plays in the perception of synthesized speech, other more subjective factors, such as preference and naturalness, are also important. Since the intelligibility of several commercially available text-to-speech systems has now reached a level of intelligibility approaching that of natural speech (Logan, Pisoni, & Greene, 1985), these somewhat more subjective factors have become topics of interest in evaluating devices that produce synthetic speech and in assessing their acceptance by users in specific applications where voice output systems may be used. Despite relatively high levels of intelligibility, synthetic speech still sounds mechanical and machine-like and users often focus on this as the major problem in adopting synthetic speech in an application using voice output. Clearly, in order to take advantage of voice output devices using synthetic speech, it will be necessary to improve their naturalness and make them sound more human-like.

In the past, researchers have assessed the perceived quality of speech produced under a number of different conditions, including vocoding and other electronic manipulations of natural speech by using multidimensional scaling techniques, isopreference methods, semantic differential scales, and other rating methods. A review of these approaches to evaluating the subjective reaction of listeners to speech after having undergone various types of processing may be found in Nusbaum, Schwab, & Pisoni (1984).

Research specifically directed at examining listeners preferences among different types of synthesized speech was included in an experiment conducted several years ago by Nye, Ingemann, & Donald (1975). They assessed the comprehension of synthetic speech produced by several different algorithms by measuring the time required to answer questions about short passages produced by each of the algorithms. Moreover, Nye et al. also examined the relationship between comprehension and listeners preference for the different speech types by presenting subjects with pairs of short passages produced by two different algorithms and determining which system listeners preferred. These results were then compared with comprehension performance using the same stimuli. Nye et al. found that listeners preferred those speech algorithms that were found to be most comprehensible in the previous tests. However, the complex nature of the factors related to comprehension (segmental intelligibility, word recognition, and prosody, for example) make it difficult to determine the degree to which each component contributed to the overall results. In other words, it may have been more fruitful to examine one of the factors contributing to comprehension, such as intelligibility, and determine its relation to preference among different synthetic speech algorithms.

In another study, McHugh (1976) also investigated the relationship between preference and comprehension of different synthetic voices. Listeners were presented with several different types of synthetic speech which differed in stress. Their task in this experiment was to decide the extent to which the different synthetic voices sounded "good" or "bad" using a rating scale. In a subsequent experiment, McHugh used the same synthetic voices to generate passages that were presented to subjects followed by questions to assess comprehension. She obtained a similar pattern of results in both the rating and comprehension tasks, suggesting that comprehension and preference were closely interrelated. Specifically, the preferred voices were also the more comprehensible. Thus, a similar pattern of results was obtained by both McHugh (1976) and Nye, Ingemann, & Donald (1975). Performance on a comprehension task was directly related to subjective preference.

More recently, Nusbaum, Schwab, & Pisoni (1984) carried out a subjective evaluation study in which listener's preferences among natural speech and the output from two text-to-speech systems, MITalk and Votrax, were examined. Using a specially constructed evaluation questionnaire, Nusbaum et al. found that subjects judged synthetic speech to be more coarse, rough sounding, and harsh than natural speech. Subjective differences between the two synthesized voices tended to follow intelligibility: MITalk was rated more positively than Votrax on adjective pairs such as hard/easy, gentle/harsh, and halting/fluent which corresponded to the results obtained using tests of segmental intelligibility which showed MITalk to be more intelligible than Votrax.

The results of the Nusbaum et al. (1984) study suggested that intelligibility was an important component of preference but the results did not permit a more precise elaboration of the nature of the relationship between preference and intelligibility or any of the other factors that presumably underlie subjective preference. Since the intelligibility differences between MITalk and Votrax were substantial, subject's awareness and attention to other qualities differentiating the synthesized voices may have been obscured. In other words, subjects may have used gross differences in intelligibility alone when they made their evaluations. Also, the use of adjective pairs in the rating scales may have biased subjects judgements by providing them with somewhat artificial dimensions on which to base their decisions.

The issues raised by the results of the Nye et al. (1975), McHugh (1976), and Nusbaum et al. (1984) studies led us to the following question: What would subjects do if they were given output from several synthesizers of similar intelligibility and were asked to make direct preference judgements using whatever criteria they thought was important? The present experiments were designed to address this question.

Subjects were presented with the same sentence generated by two different text-to-speech systems in an A/B format. Their task was to decide which of the two voices they preferred and then make a confidence judgement on their decision. The problems associated with the use of adjective pairs to assess preference were addressed in the present experiment by the use of an A/B paired comparison method. In using this procedure, only those criteria the subjects themselves considered relevant determined which of the two voices they preferred. Moreover, the A/B paired comparison test provides an objective, forced-choice measure of the observers preference for one synthetic voice over another.

Experiment I

The first experiment served to validate the A/B paired comparison procedure as well as provide a baseline measure of the relation between preference and intelligibility. Based on previous tests of segmental intelligibility carried out in our laboratory using the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965), output from three text-to-speech systems was chosen for use in the first experiment: DECtalk, Prose, and MITalk. These tests yielded error rates of 3.25% for DECtalk, 5.72% for

Prose, and 7.0% for MITalk (see Logan et al., 1985). The stimuli chosen for use in the first experiment were Harvard Psychoacoustic sentences (Egan, 1948). One reason for the choice of these particular stimuli was that sentences enabled the prosodic features of the text-to-speech systems to be evaluated in some gross manner. If individual words were used as stimuli the effects of prosody would be minimal. Furthermore, these sentences were grammatical and easily understandable, making the subject's task less dependent on intelligibility alone when making preference decisions.

Method

Subjects. A total of 38 native speakers of English participated as paid observers. All reported no history of a speech or hearing disorder. None had any extensive experience listening to synthetic speech prior to this experiment.

Stimuli. The stimuli consisted of 40 Harvard Psychoacoustic sentences generated using the DECtalk 2.0, Prose 2000 V3, and MITalk-79 text-to-speech systems. The output of each synthesizer was recorded on audio tape, low pass filtered at 4.8 kHz, and then digitized at 10 kHz using an A/D converter with 12-bit resolution. Each sentence was then edited into individual waveform files using WAVES, a waveform editing program developed in our laboratory (see Luce & Carrell, 1981).

Procedure. The sequence of events during each experimental trial was as follows: Subjects listened to the sentence produced by Voice A, and then, after a 500 ms interval, they listened to the same sentence produced by Voice B. After hearing the pair of sentences, subjects decided which voice they thought was the most natural-sounding in the pair by pressing either "A" or "B" on their response boxes. After subjects recorded their preference judgement, they were also required to indicate how confident they felt about their decision on a seven point rating scale that ranged from "just guessing" to "very confident". After all subjects had recorded their rating response, the next trial began. The order of presentation for each voice was randomized across trials.

At the beginning of each experimental session, subjects were told that they would be participating in an experiment in which they would hear synthetic speech produced by a computer. Following a description of the sequence of events in each trial, subjects were told to treat each trial independently of the preceeding trials. This was done in order to discourage the use of a strategy in which subjects would always choose the same voice regardless of the context. Subjects were also told that in this experiment there were no right or wrong answers, and that we were interested only in their relative judgements about which voice they preferred. After receiving instructions, subjects were given four practice trials followed by forty experimental trials. Each experimental session lasted about 30 minutes.

Subjects were tested in groups of three to six in a quiet room. Stimuli were presented at 80 dB SPL against a background of 50 dB white noise over TDH-39 matched and calibrated headphones. The white noise was generated using a Grason-Stadler 1724 noise generator. A PDP 11/34 computer was used to control the presentation of stimuli and to record subjects responses. Three experimental conditions resulted from each pair-wise comparison between the synthesizers: 1) DECtalk vs. Prose, 2) DECtalk vs. MITalk, and 3) Prose vs. MITalk. Three measures were obtained from subjects on each trial: 1) a forced-choice preference response, 2) a response latency for the preference decision, and 3) a confidence rating for that decision.

Results and Discussion

The overall proportion of preference responses for each voice is shown in Figure 1.

Insert Figure 1 about here

The largest difference in the pair-wise comparisons was between DECtalk and MITalk, followed by DECtalk and Prose, and then Prose and MITalk. In each pair-wise comparison, the preferred voice was also the more intelligible voice: DECtalk over Prose, DECtalk over MITalk, and Prose over MITalk. All differences in preference were statistically reliable except for the Prose/MITalk comparison (see description of analyses below).

Insert Figure 2 about here

The mean confidence rating for the subjects' preference decisions in each of the comparison conditions is shown in Figure 2. As in the overall preference data, the mean ratings also follow the rank ordering of segmental intelligibility. That is, confidence ratings for DECtalk were higher than the confidence ratings for Prose and, in turn, the confidence ratings for DECtalk were higher than the confidence ratings for MITalk. Finally, the confidence ratings for Prose were higher than the confidence ratings for MITalk. These differences were all statistically reliable (see description of analyses below).

Insert Figure 3 about here

A more detailed presentation of the data is shown in Figure 3. This figure shows the distribution of the preference responses as a function of the confidence ratings in the DECtalk/Prose comparison (top panel), MITalk/DECtalk comparison (middle panel), and the Prose/MITalk comparison (bottom panel). In the DECtalk/Prose comparison, subjects displayed a high degree of confidence in their preference decisions as shown by the higher proportion of responses towards the "most confident" end of the rating continuum compared to the "just guessing' end of the rating continuum. This panel also shows that when subjects were more confident of their responses, the differences in preference response between the pair of synthetic voices were greater. An analysis of variance was performed on the data from the DECtalk/Prose condition and the results revealed significant main effects for voice [$F(1, 15)=18.92$, $p<0.0006$] and confidence rating category [$F(6, 90)=5.35$, $p<0.0001$]. In addition, a significant interaction between voice and confidence rating category was also obtained [$F(6, 90)=3.74$, $p<0.0023$]. Thus, DECtalk is preferred over Prose and subjects tended to be more confident of their preference decision when they

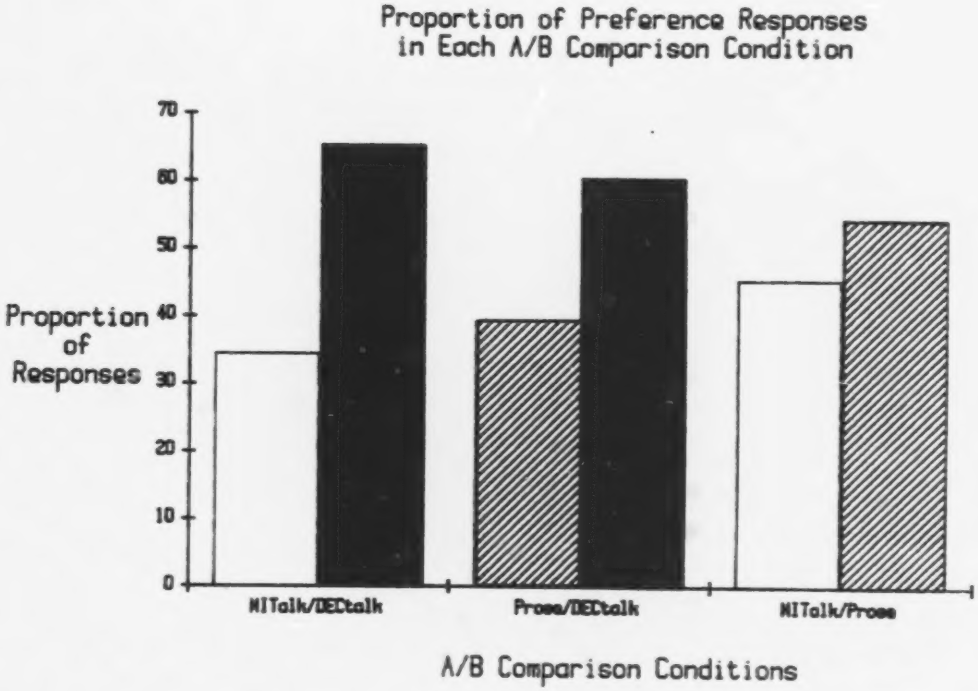


Figure 1. Overall proportion of preference responses for each voice in each comparison condition (Experiment I).

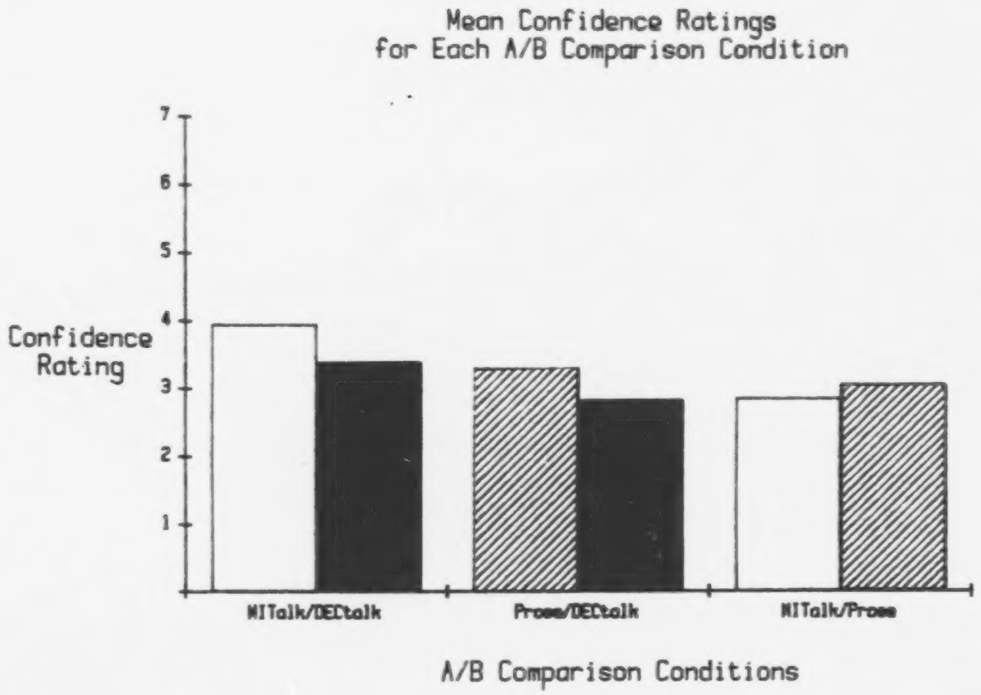


Figure 2. Mean confidence rating for preference decision in each comparison condition (Experiment I).

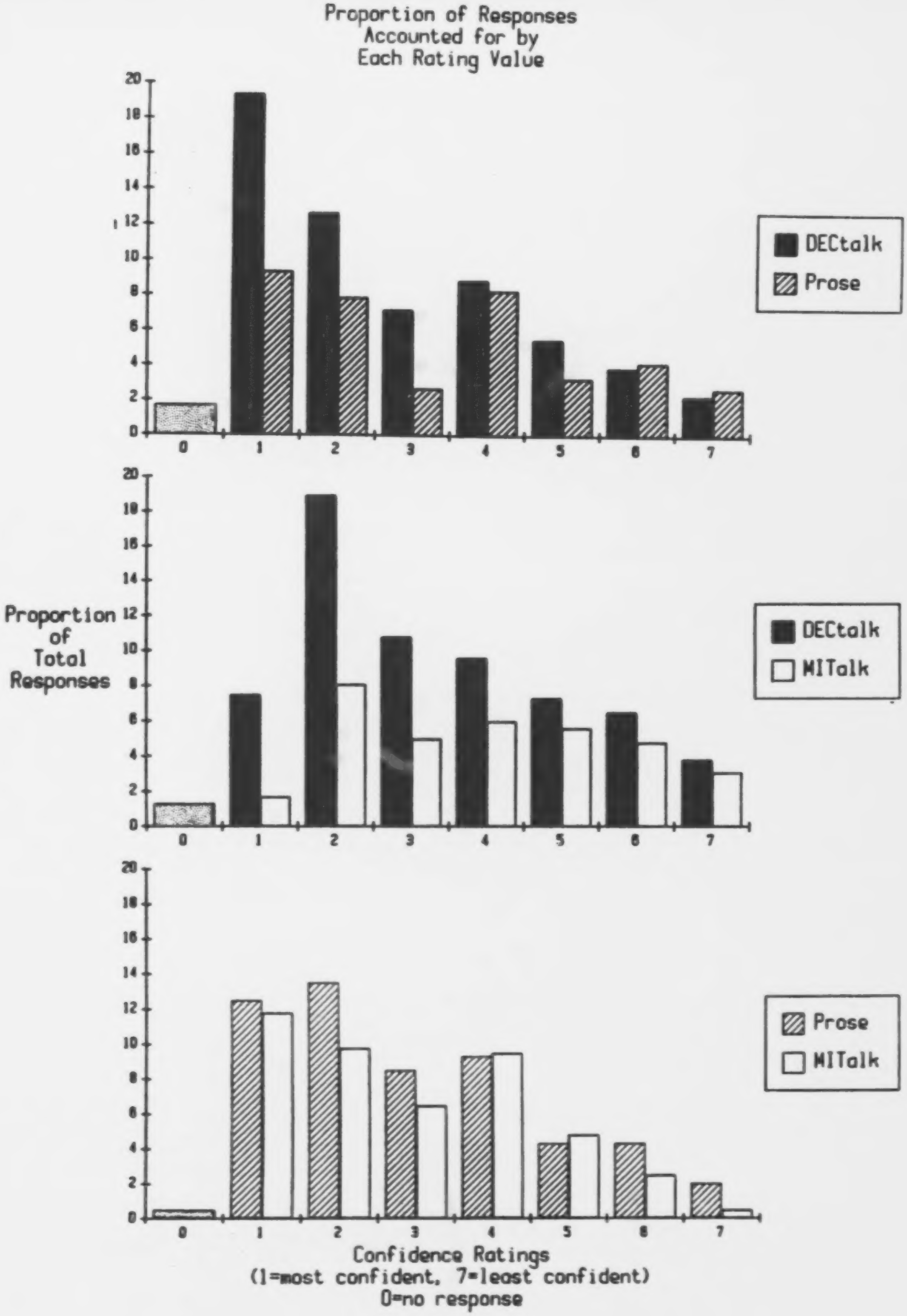


Figure 3. Proportion of preference responses for each confidence rating category in the DECtalk/Prose comparison condition (top panel), the DECtalk/MITalk condition (middle panel), and the Prose/MITalk condition (bottom panel) (Experiment I).

chose DECtalk than when they preferred Prose.

The distribution of preference responses for each rating category in the DECtalk/MITalk comparison is shown in the middle panel of Figure 3. A pattern of results similar to that found in the DECtalk/Prose comparison may be observed: a greater proportion of preference responses were assigned to the "most confident" end of the continuum and differences in preference between voices became greater as the confidence ratings increased. An ANOVA was carried out to assess the reliability of these effects and significant main effects for voice [$F(1, 11)=54.35$, $p<0.00001$] and confidence rating category [$F(6, 66)=5.31$, $p<0.0002$] were obtained. In addition, a significant interaction between voice and confidence rating category was also observed [$F(6, 66)=3.65$, $p<0.0034$]. Overall, subjects preferred DECtalk over MITalk; also, they tended to be most confident of their preference decision when they chose DECtalk. The only anomolous point is the overall attenuation of responses in the "most confident" rating category; however, the difference between the two voices is still present even in this rating category.

The distribution of preference responses for each rating category in the Prose/MITalk comparison is shown in the bottom panel of Figure 3. In this condition, the tendency for a greater proportion of responses for the "most confident" end of the continuum was still present but no strong voice preference emerged. An ANOVA was performed on these data and the results indicated that this was the only comparison condition in which a significant effect for voice failed to emerge [$F(1, 9)=2.44$, $p<0.1524$]. There was, however, a significant main effect for confidence rating category [$F(6, 54)=3.21$, $p<0.0091$]. No significant interaction between voice and confidence rating category was obtained [$F(6, 54)=0.59$, $p<0.7347$]. Apparently, subjects were equivocal in their preferences between MITalk and Prose.

A further set of analyses were carried out to determine the relationship, if any, between response times for the initial A/B preference decision and the subsequent confidence ratings. The assumption underlying these analyses was that a fast response time for the preference decision would reflect a greater degree of confidence in that decision. Similarly, a slow response time for the preference decision would indicate much greater uncertainty in deciding which voice was preferred. If the response time measure and the confidence rating for the preference decision were both measuring the same degree of certainty in the decision, these two measures should be positively correlated. Separate Pearson product moment correlations were calculated for each of the three conditions in the experiment. Each took the form of a partial correlation in which the effect of which voice was preferred was removed, leaving only the effect of response time and confidence rating (The differences between the partial correlations and the correlations with the effect of voice present were minimal). The partial correlation coefficients for each condition were as follows: DECtalk/Prose condition, $r=+0.302$; Prose/MITalk condition, $r=+0.264$; MITalk/DECtalk condition, $r=+0.460$.

These correlations show a small positive relationship between response times for the A/B preference decision and the subsequent confidence ratings, indicating that the confidence ratings are reasonably valid indicators of subjects uncertainty in their preference decision. The small size of the correlation coefficients may have been due to the post hoc nature of the confidence rating response within each trial: enough time may have elapsed after the initial preference decision to make the confidence ratings a less effective indicator of certainty than if it were possible for subjects to indicate preference and confidence rating within the same response. Despite the small size of the correlation coefficients for confidence ratings and response times, the size of the coefficients does correspond to the preference judgements: the two highest coefficients were obtained in the DECtalk/Prose and MITalk/DECtalk comparison conditions which were also the two conditions in which one voice was preferred over the other for a significant proportion of the responses. Thus, it appears that when subjects showed a definite preference for one voice over the other, response times and confidence ratings tended to be reasonably correlated.

A final analysis was carried out in order to determine if any biases existed in the button press response when making the A/B preference decision. Since the order of Voice 1 and Voice 2 varied randomly from trial to trial, subjects ideally should have shown no preference for one response over the other, within the limits imposed by chance. This analysis indicated a strong preference for pressing the right-hand button. However, this effect varied as a function of the confidence subjects had in their responses. When subjects were sure of their response, they tended to show no right-hand button bias. But, when subjects were unsure of their preference decision, a right-hand button bias began to emerge in their responses. Because virtually all of our subjects were right-handed (86.8% of subjects reported they were right-handed, 7.9% were left-handed, and 5.3% were ambidextrous), this response bias can be attributed to a handedness preference. It is important to keep in mind that despite the presence of a response bias, the voice preference results described above were quite systematic. The overall effect of this bias was minimal and only occured when subjects were unsure of their preference decision which for most subjects was only for a small proportion of the trials. The relationship between a right button-pressing bias and confidence ratings is displayed in Figure 4. When subjects are more confident in their preference decision, the response bias becomes more attenuated.

Insert Figure 4 about here

In summary, several general conclusions may be drawn from the results of this initial experiment using the A/B preference task. First, intelligibility appears to play a major role in determining a subject's preference for one synthetic voice over another. Results from two of the three comparison conditions indicated that the most preferred synthetic voice could be predicted from the segmental intelligibility scores of the synthetic voices. In short, subjects preferred to listen to the more intelligible of two synthetic voices, even when differences in intelligibility were small, as in the DECtalk/Prose comparison. Of course, to determine more precisely the role of intelligibility in preference decisions is difficult in the context of the present experiment alone. A further discusion of this point will be found below in the general discussion.

Our experiment also demonstrated the usefulness of the A/B method of paired comparisons as a means of examining differences in the subjective quality of synthetic speech produced by rule. The strength of the conclusions concerning the relation between intelligibility and preference is further supported by the convergence of the A/B preference judgement, response latency, and confidence rating data. The finding that consistent and systematic data was obtainable using this procedure suggested that further investigations were clearly warranted. To this end, a second experiment using the A/B procedure was undertaken. In this second experiment, a new set of

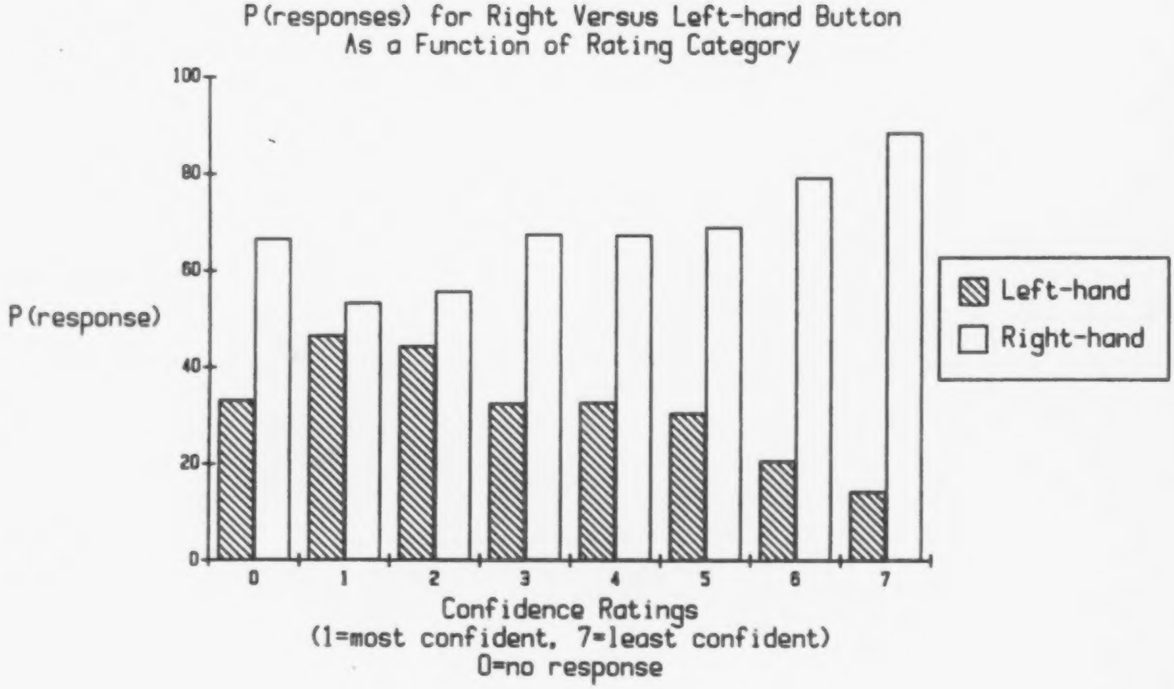


Figure 4. Proportion of responses in each confidence rating category as a function of the left/right response button (Experiment I).

stimulus materials were used. Instead of the Harvard sentences, the Phoneme Specific Sentences (PSS) developed by Huggins et al. (1985) were used and some attempt was made to relate preference and rating data to specific phoneme classes.

Experiment II

The robustness of the findings obtained in Experiment I suggested that other sets of stimuli used in the same A/B paired comparison procedure might also provide useful information. We chose the PSS (Huggins et al., 1985) because they were specifically developed for use in evaluating speech processing devices. Each sentence was "loaded" with phonemes of a particular class: for example, several sentences contained a large proportion of fricatives, while others contained a large proportion of voiced stops, etc. Altogether, a total of 92 sentences were used. These sentences are listed in Appendix 1. By examining the relationship between the preference for specific sentences and the classes of phonemes contained within those sentences, some understanding of the perceptual basis for preference associated with different classes of phonemes produced by the different synthetic voices can be obtained. These data should therefore provide a greater body of knowledge regarding what acoustic-phonetic and phonological parameters affect preference in evaluating different types of synthetic speech produced by rule.

The choice of voices to be used in this experiment was based, in part, on MRT performance data. We wanted to use the same voices as in the first experiment in order to make the difference in the test sentences the only variable that changed between the first and second experiments; however, since output from the MITalk text-to-speech system was no longer available, we chose output from the Infovox system instead. The reason for the choice of the Infovox system was its similarity to the MITalk system in terms of overall intelligibility (see Logan et al., 1985).

Because the PSS stimuli have not been widely used prior to this experiment, we also decided to collect data on the intelligibility of these materials. These data were collected in an independent experiment and are reported in a companion paper (see Logan & Pisoni, 1986). Each of the voices used in the present experiment were tested in a simple identification/transcription task in which subjects were presented with all 92 sentences in a random order. The subjects task was to transcribe what he or she heard. The transcriptions were scored for exact matches to the intended output and results, reported as percent correct, were as follows: DECtalk - 56.34%; Prose - 56.24%, and Infovox - 24.91%. Using the MRT, the differences in performance between DECtalk and Prose were somewhat larger while the differences between Prose and Infovox were substantially smaller. However, the rank ordering of the intelligibility scores of the voices using the PSS stimuli remained identical to that obtained with isolated words using the MRT and the Harvard sentences. Table 1 shows the results obtained using the voices studied in the present investigation with several different kinds of test materials.

Insert Table 1 about here

Table 1

Measure of Intelligibility for Several

Text-to-Speech Systems
(Percent Error)

	Measure of Intelligibility		
Text-to-Speech System	MRT (closed/open)	Harvard Sentences	Phoneme Specific Sentences
DECtalk	3.3/13.3	4.7	43.7
Prose 2000	5.7/19.9	[6.5*]	43.8
MITalk	6.9/24.6	6.7	-
Infovox	12.5/37.2	-	75.1

*This error rate was obtained using an earlier version of the Prose 2000 than used in the MRT and PSS evaluations.

Method

Subjects. Thirty subjects were obtained from a subject pool of students taking an introductory psychology course at Indiana University. Subjects received course credit for their participation. All subjects reported no history of a speech or hearing disorder and none of the subjects had extensive experience with sythetic speech prior to the present experiment.

Stimuli. The stimuli used in the second experiment were the Phoneme Specific Sentences (PSS) constructed by Huggins et al. (1985) to examine the effect of different methods of speech processing on the perception of specific classes of phonemes. A total of 92 sentences, each containing a large proportion of phonemes from a particular phoneme class, were generated using the DECtalk, Prose 2000, and Infovox SA101 text-to-speech systems. As in the first experiment, the stimuli were digitized and edited into individual sentences. To ensure compatability among the stimuli in terms of overall signal level, the sentences were processed using a waveform modification program (Bernacki, 1981). A target value of 50 dB RMS was chosen to minimize clipping and yet provide an adequate signal level.

Procedure. The same procedure used in the first experiment was also used in this study. Each experimental session lasted approximately 45 minutes.

Results and Discussion

The overall proportion of preference responses for each voice is shown in Figure 5.

Insert Figure 5 about here

A pattern of results similar to that obtained in Experiment I was found: the largest differences in preference correspond to the largest differences in intelligibility between the voices. However, in this experiment, in all three comparison conditions the differences between the proportion of preference responses for each voice were significantly different (see description of analyses below).

The mean confidence ratings for the preference decisions are shown in Figure 6.

Insert Figure 6 about here

Statistically reliable differences were obtained in the mean confidence ratings for the DECtalk/Prose and DEC/Infovox comparisons. No significant difference in the mean confidence ratings was observed in the Infovox/Prose comparison (see description of analyses below).

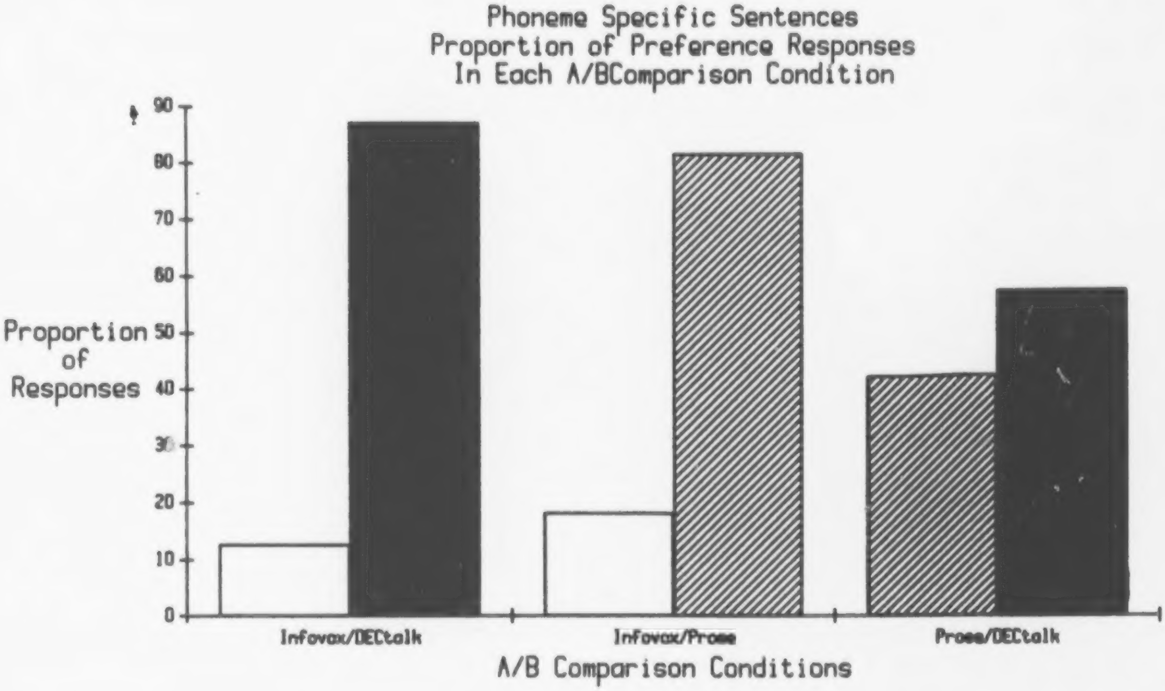


Figure 5. Overall proportion of preference responses for each voice in each comparison condition (Experiment II).

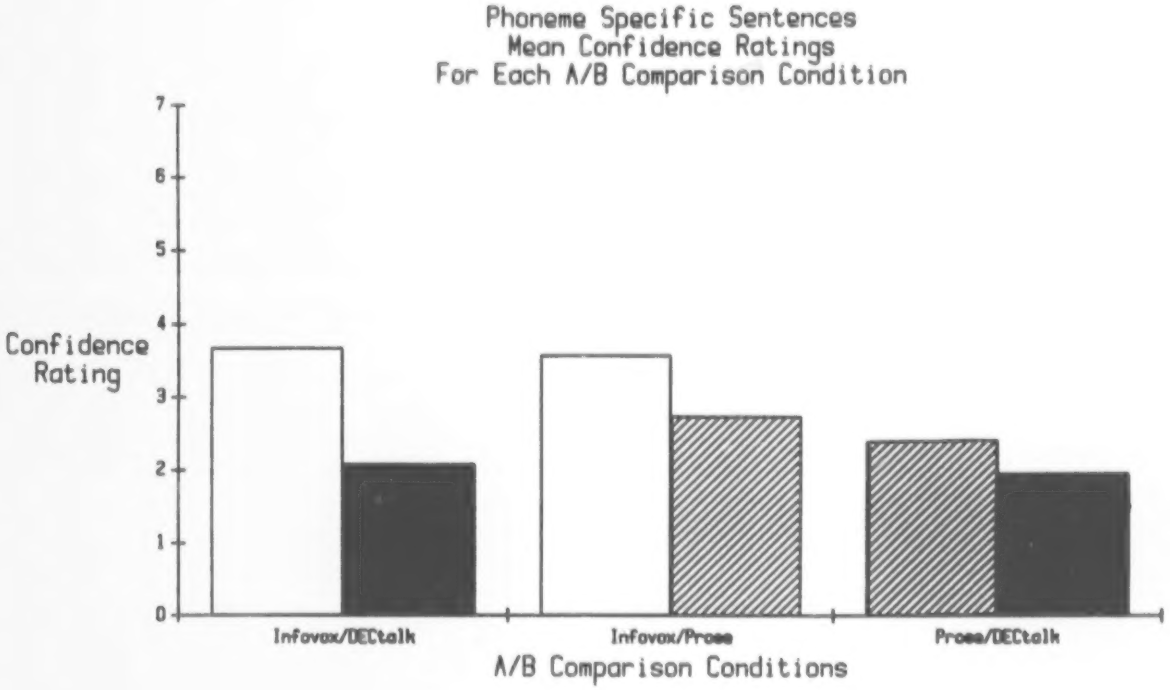


Figure 6. Mean confidence ratings for the preference decision in each comparison condition (Experiment II).

Figure 7 shows the preference responses as a function of rating category for the DECtalk/Prose comparison (top panel), the DECtalk/Infovox comparison (middle panel), and the Infovox/Prose comparison (bottom panel).

Insert Figure 7 about here

As in Experiment I, subjects preferred DECtalk over Prose by a large margin in the two end-point rating categories at the "most confident" end of the rating scale. When subjects were more uncertain of their preference decision, DECtalk was still preferred but the difference between the proportion of preference responses for each voice was much smaller. In general, subjects were confident of their preference decisions as indicated by the large proportion of responses at the "most confident" end of the rating scale for both voices. This description of the data from the DECalk/Prose comparison was supported by an ANOVA comparing the proportion of preference judgements for each voice in each rating category. Significant main effects for voice $[F(1, 9)=6.66, p<0.0297]$ and confidence rating category $[F(6, 54)=3.09, p<0.0114]$ were obtained, but no significant interaction between voice and confidence rating category $[F(6, 54)=1.46, p<0.2101]$ was found.

The preference responses for the DECtalk/Infovox condition as a function of rating category are shown in the middle panel of Figure 7. In every rating category, DECtalk was consistently prefered over Infovox. The difference is especially noticeable at the "most confident" end of the rating scale but the trend is quite noticable over the entire range of rating responses. An ANOVA revealed significant main effects for voice $[F(1, 9)=274.2, p<0.00001]$ and rating category $[F(6, 54)=4.16, p<0.0017]$ and an interaction of voice and rating category $[F(6, 54)=6.09, p<0.0001]$, confirming the trends shown in Figure 7.

The preference data for the Prose/Infovox condition as a function of rating category is shown in the bottom panel of Figure 7. In all rating categories, Prose was preferred over Infovox. The largest difference was observed between the two voices at the "most confident" end of the rating scale, suggesting that subjects were more confident in their choice decision if they chose Prose over Infovox. An ANOVA on these data revealed a significant main effect for voice $[F(1, 9)=76.36, p<0.00001]$ but no significant main effect for confidence rating category $[F(6, 54)=1.16, p<0.3425]$ nor any significant interaction between voice and confidence rating $[F(6, 54)=2.0, p<0.0769]$. Despite the appearance of a trend for the least preferred voice, Infovox, to have a larger proportion of responses at the "least confident" end of the rating scale, the effect was not substantial enough to produce a significant interaction between voice and confidence rating.

As in Experiment I, several analyses were carried out to examine the relationship between response time for the preference judgements and confidence ratings. The partial correlation coefficients (removing the effect of voice) for each comparison condition were as follows: DECtalk/Prose, $r=+0.208$; DECtalk/Infovox, $r=+0.120$; and Infovox/Prose, $r=+0.206$. All of these correlations were lower than any of those obtained in the first experiment. Also, little difference existed between the partial correlations reported here and the corresponding correlations with the effect of voice present.

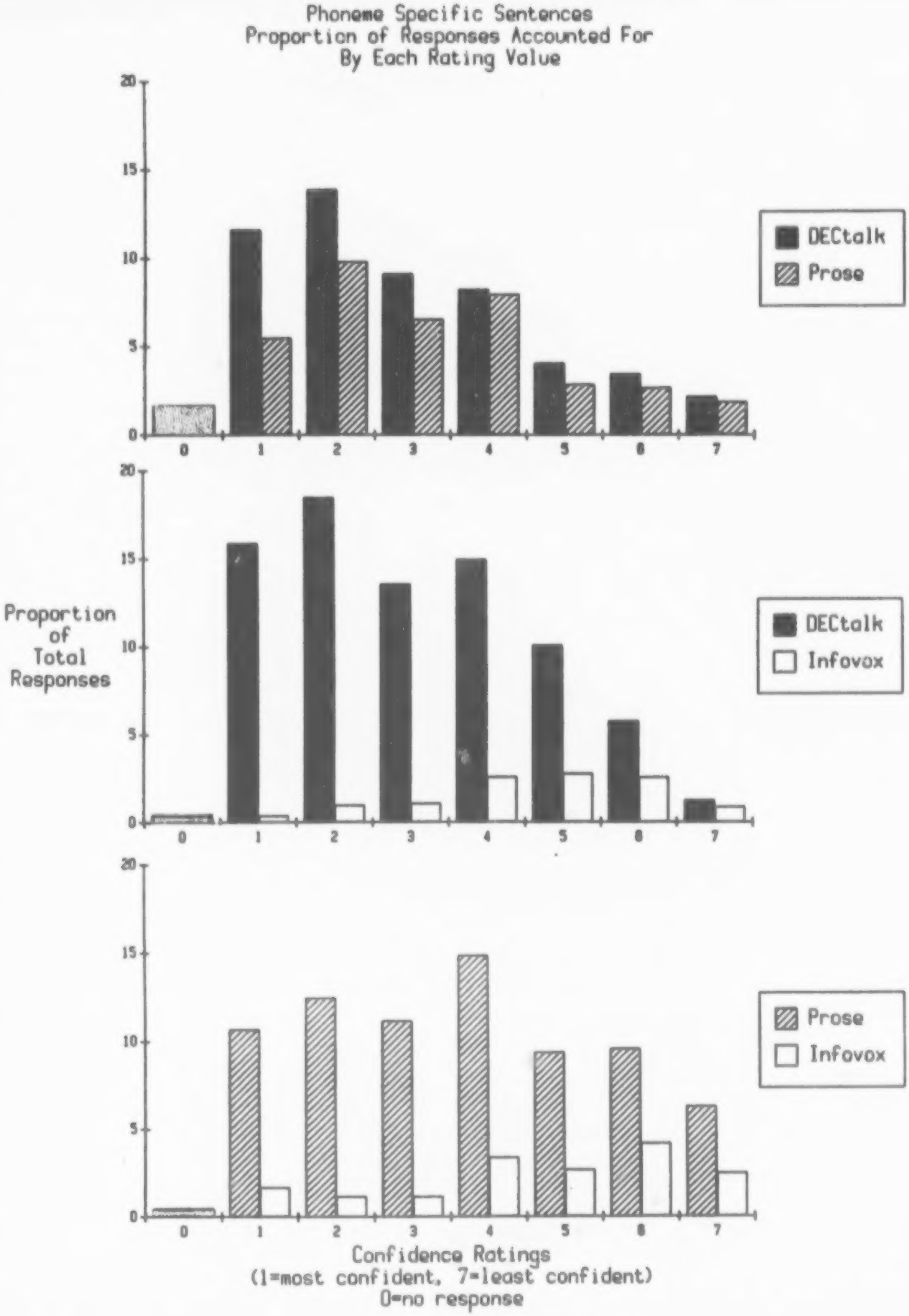


Figure 7. Proportion of preference responses for each confidence rating category in the DECtalk/Prose condition (top panel), the DECtalk/Infovox condition (middle panel), and the Infovox/Prose condition (bottom panel) (Experiment II)

The consistently low correlations obtained in both Experiments I and II between response time and confidence rating may be an indicator of the complexity of the task facing subjects. The temporal separation between the two measures in each trial may be sufficient to make the second measure, the confidence rating, more dependent on factors such as memory and post-perceptual decision making. Another aspect of the task that may affect the performance of subjects is the two different sets of stimuli used in Experiments I and II. Substantially smaller correlations between response times and confidence ratings were obtained with the PSS stimuli than with the Harvard sentences. This result may have been related to the somewhat anomalous nature of many of the PSS stimuli. In order to achieve the desired loading of phonemes from a given phonetic class into a single sentence, many of the resulting sentences were often constructed using unusual combinations of words that rarely occur in English. These constraints were reflected in the relatively low scores in transcription task used to assess the intelligibility of these phoneme specific sentences; the comparible transcription data for the Harvard sentences yielded substantially greater levels of intelligibility when scored using the same criteria (see Table 1).

A final set of analyses was carried out to study the relationship between the different phonetic classes that served as the basis for the creation of the PSS stimuli and the proportion of preference responses for each voice in each pair-wise comparison. Each analysis took the form of an ANOVA in which the two factors were voice (2 voices) and phonetic class (18 categories). In the Prose/DECtalk comparison, a significant main effect for voice was obtained [$F(1, 9)=9.26$, $p<0.05$], as well as a significant interaction between voice and phonetic category [$F(17, 153)=183.4$, $p<0.05$]. The absence of a main effect for phonetic category and the presence of an interaction between voice and phonetic category indicates that preference changed as a function of phonetic category although overall, DECtalk was preferred over Prose. In the DECtalk/Infovox comparison, a significant main effect for voice was obtained [$F(1, 9)=331.69$, $p<0.0001$], but no effect of phonetic category or interaction between voice and phonetic category was found. This analysis indicated that DECtalk was strongly preferred over Infovox regardless of phonetic category. In the Infovox/Prose comparison, a significant main effect for voice was obtained [$F(1, 9)=79.53$, $p<0.0001$], as well as a significant interaction between voice and phonetic category [$F(17, 153)=3.47$, $p<0.0001$]. The absence of a main effect for phonetic category and the presence of an interaction between voice and phonetic category suggests that although Prose was preferred over Infovox, the degree to which this relationship was observed changed as a function of the phonetic category that was represented by the sentences. Thus, when subjects are forced to make fine phonetic distinctions they have to attend closely to the specific phoneme classes and not only the overall quality of the voice. The mean proportion of preference responses for each voice in each phonetic category is shown in Table 2.

Insert Table 2 about here

In summary, the preference data obtained using the PSS stimuli was, for the most part, quite similar to the data obtained using the Harvard sentences. Again, small differences in segmental intelligibility were positively correlated with preference. Thus, the major findings of Experiment I were replicated in Experiment II using different sentence materials and another

Table 2

Proportion of Preference Responses for Phoneme Specific Sentence Categories

PSS Categories	Proportion of Preference Responses		
	DECtalk/Prose	DECtalk/Infovox	Infovox/Prose
1) all fricatives	55.0/45.0	100.0/00.0	6.7/93.3
2) all stops & affricates	63.3/36.7	93.3/6.7	16.7/83.3
3) all consonant phonemes	49.3/50.7	96.7/3.3	5.0/95.0
4) glides except l & vowels	40.0/60.0	85.0/15.0	5.0/95.0
5) glides	60.0/40.0	95.0/5.0	0.0/100.0
6) glides & nasals	45.5/54.5	85.9/14.1	12.0/88.0
7) all labials	50.0/50.0	85.0/15.0	5.0/95.0
8) nasals	56.3/43.7	90.0/10.0	12.9/87.1
9) nasals + l	75.0/25.0	90.0/10.0	15.0/85.0
10) all tongue tip	50.0/50.0	82.9/17.1	15.0/85.0
11) all unvoiced consonants	48.0/52.0	84.0/16.0	6.0/94.0
12) unvoiced fricatives	75.0/25.0	100.0/00.0	15.0/85.0
13) unvoiced stops	71.5/28.5	76.0/24.0	48.0/52.0
14) unvoiced stops & affricate	73.3/26.7	82.0/18.0	31.0/69.0
15) voiced fricatives	75.0/25.0	87.5/12.5	36.7/63.3
16) all voiced consonants	71.5/28.5	86.0/14.0	36.0/64.0
17) voiced stops	60.0/40.0	91.3/8.7	25.7/74.3
18) voiced stops & affricate	75.0/25.0	90.0/10.0	20.0/80.0

synthetic voice. The use of the PSS stimuli enabled a preliminary examination of the relationship between preference and individual classes of phonemes. In two of the three comparison conditions examined in this experiment, preference for one synthetic voice over another does change as a function of phonemic class under consideration. Other measures, such as the strength of the relationship between confidence ratings and response times for the preference decision, varied between the first and the second experiment. However, these differences appeared to be related to an interaction between the PSS stimuli and the task requirements in the second experiment.

General Discussion

The results of Experiment I and Experiment II are very similar and appear to be related to two factors. First, the A/B paired comparison task itself clearly yields useful and reliable information concerning the preferences of inexperienced subjects listening to several different kinds of synthetic speech produced by rule. Although certain aspects of the task, such as the confidence ratings, did not display as strong a relationship to other measures as might be expected, the overall task provided systematic data that proved useful in revealing differences in preference among several types of synthetic speech studied here.

Secondly, the results suggested that segmental intelligibility plays an important role in determining subjects' preference for one voice over another. Even when differences in intelligibility were small, preference was positively correlated with intelligibility. However, knowledge of this relationship between intelligibilty and preference must be viewed as incomplete. To determine the relative contribution of intelligibility to preference judgements requires the comparison of two voices of equivalent intelligibility. Further experiments investigating the effects of manipulating different parameters within the same synthetic voice and their effects on preference are underway in our laboratory.

The results of the present experiments suggest several additional extensions of our initial appraoch to studying preferences among synthetic voices. Further research is currently underway to examine the relationship between this method of assessing preference and other more subjective methods such as questionnaires and rating scales (see Nusaum et al., 1984). Also, the relationship between intelligibility and preference warrants further examination using other types of stimuli such as words, sentences, and even longer passages of connected fluent speech produced by rule. In addition, the relationship between preference and naturalness is also of interest. It may not neccesarily be the case that these two subjective attributes are the same although they are likely to be interrelated in some complex way.

Finally, we feel that the information obtained in this study has a number of implications for further work involving "real world" applications where preference may be directly related to user acceptance of voice output devices utilizing synthetic speech produced by rule. At the present time, almost all synthetic speech produced by rule, even highly intelligible synthetic speech such as that produced by DECtalk and Prose, sounds mechanical and unnatural to most human observers who are not speech scientists. As more and more effort is devoted to research on the acoustic cues to naturalness, comparable efforts will also need to be devoted to developing new methods and techniques to assess these more subjective qualities and to determine the relationships between segmental intelligibility, subjective preferences, and naturalness. In some applications of speech synthesis, there may be some advantages to having an unnatural and very mechanical sounding voice. But in other applications, naturalness and user acceptability may be extremely desirable attributes that listeners will come to expect in voice output devices using synthetic speech.

References

Bernacki, R. (1981). WAVMOD: A program to modify digital waveforms. In *Research on Speech Perception. Progress Report No. 7*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Egan, J. (1948). Articulation testing methods. *Laryngoscope*, *58*, 955-991.

Huggins, A. F., & Nickerson, R. S. (1985). Speech quality evaluation using "phoneme-specific" sentences. *Journal of the Acoustical Society of America*, *77*, 1896-1906.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, *37*, 158-166.

Logan, J.S., Pisoni, D. B., & Greene, B. G. (1985). Measuring the segmental intelligibillity of synthetic speech: Results from eight text-to-speech systems. *Research on Speech Perception. Progress Report No. 11*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Carrell, T. D. (1981). Creating and editing waveforms using WAVES. In *Research on Speech Perception. Progress Report No. 7*, Bloomington, IN: Speech Research Laboratory, Indiana University.

McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. *Naval Research Laboratory Report 8015*.

Nusbaum, H. C., Schwab, E. C., & Pisoni, D. B. (1984). Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability. *Research on Speech Perception. Progress Report No. 10*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Nye, P. W., Ingemann, F., & Donald, L. (1975). Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. *Haskins Laboratories: Status Report on Speech Perception SR-41*.

Appendix 1

Phoneme Specific Sentences
(from Huggins & Nickerson, 1985)

Fricatives -
His vicious father had seizures.
Whose shaver has three fuses?
Three of the chefs saw the thieves.

Stops and affricates -
Which tea party did Judge Baker go to?
We'd better buy a bigger dog.
Georgie had to chew tobacco.

Consonants -
If the treasure vans got so much publicity we think you should hide your share.
The voyagers have ground the crankshaft with unimpeachable precision.
The old-fashioned jacket was giving you both so much humourous pleasure.
The average disillusioned gambler thinks he wishes for a cheap yacht.
Nothing could be further from reality than his illusion of chasing your gorgeous sheep away.
She thinks even the pale rouge you bought was much too gaudy for her age.

Glides except l -
Why were you away a year Roy?
Why were you weary?

Glides -
Our lawyer will allow your rule.
Our rule will allow you a lawyer.
We really will allow you a ruler.

Glides and nasals -
You were wrong all along.
I know you're all alone.
When will our yellow lion roar?
An alarm rang a warning in only one room.
A lawyer may well allow a new ruling.
I'm learning my new role.
I'll remain in my narrow room.
Anyone may rely on a mailman.
I'm wearing my maroon ring.
We'll allow you a new loan.
I'll lie in an alarming manner.
Why lie when you know I'm your lawyer?
A normal animal will run away.
Mail me an aluminum railing.
I'll willingly marry Marilyn.

Appendix 1 (continued)

Phoneme Specific Sentences

Labials -
Pay my wife by five.
Weave me a web above a poppy.
Move off my pew baby!
Weep for my baby puppy.

Nasals
Nanny may know my meaning.
I'm naming one man among many.
No one knows my name.
I know many a mean man.
I know no minimum.
Many young men owe money.
When may we know your name?

Nasals plus l -
I'm well known among men.
Nine men moaning all morning.

Tongue tip -
The judge's short decision really touched the youth.
Each decision shows the jury she lies through her yellow teeth.
Such a rash allusion to dosage teases the youth.
Seth yawns at each rash allusion to the dosage.
The designers really earned the judge's derision this year.
Each allusion to Daisy's agility lessens her attention.
Each decision shows that he lies through his yellow stained teeth.
John drowned his sorrows in gin and orange juice.

Unvoiced consonants -
She swiftly passed a health check.
He steps off a path to cash a check.
I hope she chased her fox to earth.
A thick-set officer pitched out her hash.
He checked through fifty ships.

Unvoiced fricatives -
A thief saw a fish.
I saw three fish.
Three chefs face a thief.

Phoneme Specific Sentences

Unvoiced stops -
Take a copy to Pete.
Pat talked to Kitty.
Quite a cute act.
Peter took out a potato.
Kate typed a paper.

Unvoiced stops and affricates -
Chip took a picture.
A teacher patched it up.
Chat quietly to teacher.
Quite quiet at church.
Catch a paper cup.
Actuate a paper copier.
A teacher taped up a packet.
Capture a cute puppy.
A teacher typed up a paper.
Katie tacked up a cute picture.

Voiced fricatives -
They use our azure vials.
There's our azure vial.
There's usually a valve.
Those waves veer over.

Voiced consonants -
Does John believe you were measuring the gun?
Your brother's vision was gradually dimming.
The regular division was led by a young major.
I gather you will be abandoning the major revisions?
The young major's evasions were growing bolder.

Voiced stops -
Bobby did a good deed.
I begged Dad to buy a dog.
Did Bobby do a good deed?
Buy Dad a bad egg.
Dad would buy a big dog.
Why did Gay buy a bad egg?
Do you abide by your bid?
Grab a doggie bag.
A greedy boy died.

Voiced stops and affricates -
Did George do a good job?
Greg adjudged Bobby dead.

II. SHORT REPORTS AND WORK-IN-PROGRESS

Auditory Perception of Complex Sounds:
Some Comparisons of Speech vs. Nonspeech Signals*

David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405 USA

---

*Preparation of this paper was supported, in part, by NIH Research Grant NS-12179 to Indiana University and, in part, by a fellowship from the James McKeen Cattell Fund. This paper will appear in Complex Sound Perception, edited by C. S. Watson and W. A. Yost, Academic Press.

Abstract

For many years, speech researchers have been interested in the differences in perception between speech and nonspeech signals. Early studies revealed marked differences in the manner in which speech sounds were discriminated suggesting two very different modes of response, a speech mode and a nonspeech mode. Recent studies using nonspeech control patterns have raised questions about these earlier interpretations and have provided the basis for explaining several phenomena in speech perception by means of more general principles of complex auditory pattern perception. This paper summarizes the philosophy behind these nonspeech comparisons and describes two recent studies, one on temporal order perception and the other on the perception of the duration of rapid spectrum changes. Both show commonalities between speech perception and the perception of complex nonspeech patterns.

Auditory Perception of Complex Sounds:

Some Comparisons of Speech vs. Nonspeech Signals

Introduction

The study of speech perception differs in several very important ways from the study of general auditory perception. First, the signals typically used to study the functioning of the auditory system have been simple, discrete and well defined mathematically. Moreover, they typically vary along only one perceptually relevant dimension. In contrast, speech sounds involve very complex spectral relations that typically vary quite rapidly as a function of time. Changes that occur in a single perceptual dimension almost always affect the perception of other attributes of the signal. Second, most of the basic research on auditory perception over the last four decades has been concerned with problems surrounding the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanisms. In the perception of complex sound patterns such as speech, the relevant mechanisms are, for the most part, quite centrally located. Moreover, while many experiments in auditory perception and sensory psychophysics have commonly focused on experimental tasks involving discrimination of both spectral and temporal properties of auditory signals, such tasks are often inappropriate for the study of more complex signals including speech. Indeed, in the case of speech perception and probably the perception of other complex auditory patterns, the relevant task for the observer is more nearly one of absolute identification rather than differential discrimination. Listeners almost always try to identify, on an absolute basis, a particular stretch of speech or try to assign some label or sequence of labels to a complex auditory pattern. Rarely, if ever, are listeners required to make fine discriminations that approach the limits of their sensory capacities.

Given the published literature on the perception of simple auditory signals, it is generally believed, at least among researchers in the field of speech perception, that a good deal of what we have learned from traditional auditory psychophysics using simple sinusoids is only marginally relevant to the study of speech perception. Perhaps some of what is currently known about speech perception might be relevant to the perception of other complex auditory patterns which have properties that are similar to speech. At the present time, there are substantial gaps in our knowledge about the perception of complex signals which contain very rapid spectral changes such as those found in speech. And, there is little if any research on the perception of complex patterns that have the typical spectral peaks and valleys that speech signals have. Finally, our knowledge and understanding of patterns containing amplitude variations like the complex temporal patterns found in speech is also quite meager at this time. Obviously, there is a lot of basic research to do.

As Pollack (1952) demonstrated over thirty years ago, speech sounds represent a class of signals that are able to transmit relatively high levels of information with only gross variations in perceptually distinctive acoustic attributes. In other words, speech is an efficient signaling system because of its ability to exploit fundamental processing strategies of the auditory system. This theme has been taken up and expanded recently by Stevens (1980) who argues that speech signals display a certain set of general properties that set them apart from other signals in the listener's auditory environment. According to Stevens, all speech signals have three general properties or attributes in common. First, the short-term power spectrum sampled at specific points in time always has "peaks" and "valleys." That is, speech signals display up and down alternations in spectrum amplitude with frequency. These peaks in the power spectrum arise from the peaks observed in the vocal tract transfer function and correspond to the formants or vocal resonances that are so prominent in vowel and vowel-like sounds. The second general property tht speech sounds display is the presence of up and down fluctuations in amplitude as a function of time. These variations in amplitude correspond to the alternation of consonants and vowels occurring in syllabic-like units roughly every 200-300 msec. Finally, the third general property that speech signals display is that the short-term spectrum changes over time. The peaks and valleys of the power spectrum change; some changes occur rapidly -- like the formant transitions of stop consonants, whereas other changes are more gradual like the formant motions of semi-vowels and diphthongs. According to Stevens (1980), speech sounds have these three general attributes and other sounds do not and it is these attributes that distinguish speech sounds from other complex nonspeech sounds.

It should also be mentioned here that in addition to some of the differences in the signal characteristics between speech and nonspeech noted above, there are also very marked differences in the manner in which speech and nonspeech signals are processed (i.e., encoded, recognized and identified) by human listeners. For the most part, research over the last thirty-five years has demonstrated that when human observers are presented with speech signals they typically respond to them as linguistic entities rather than simply as random auditory events in their environment. The set of labels used in responding to speech are intimately associated with the function of speech as a signalling system in spoken language. Thus, speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background and experience. And, a listener's performance in identifying and discriminating a particular acoustic attribute is often a consequence of the functional role this property plays in the listener's linguistic system. It is possible to get human listeners to respond to the auditory properties of speech signals with some training and the use of sensitive psychophysical procedures. But one of the fundamental differences between speech and nonspeech signals lies in the linguistic significance of the patterns to the listener and the context into which these patterns may be incorporated.

In the sections below, we briefly summarize research on the perception of complex auditory patterns that have acoustic properties that are similar to speech sounds. The results of these studies demonstrate that complex nonspeech signals may also display perceptual characteristics that were once thought to be unique to the processing of speech signals. Our findings imply, contrary to popular belief in speech perception circles, that detailed knowledge and understanding of how complex nonspeech signals are processed by the auditory system may contribute in a number of ways to a much better understanding of speech perception. The converse is also true. New knowledge concerning the acoustic correlates of speech signals and more detailed understanding of the speech perception process may also contribute to a much better understanding of the perception of complex nonspeech auditory patterns.

Voicing Perception and VOT

Interest in categorical perception has occupied the attention of speech researchers since the late 1950s. Although early studies using nonspeech control patterns failed to find similarities with the results obtained using speech signals, several more recent studies have been more successful in demonstrating comparable categorical effects. In one study, Pisoni (1977) employed a set of nonspeech tonal patterns that differed in the relative onset time of the individual components. Examples of these signals are shown in Figure 1.

Insert Figure 1 about here

A series of experiments was carried out using these patterns to study the underlying perceptual basis of voicing perception in stop consonants that differed in voice-onset time (VOT). The results of the first experiment, shown in Figure 2, provided evidence for categorical perception of these signals. The labeling functions displayed steep slopes and the discrimination functions were non-monotonic with the physical scale and displayed peaks and valleys that corresponded to changes in the labeling probabilities. Three additional experiments were carried out in this study. All of them provided additional evidence for the presence of three distinct perceptual categories along this nonspeech stimulus continuum which were separated by narrow regions of high discriminability.

Insert Figure 2 about here

Based on these findings using nonspeech patterns that differed in relative onset time, a general account of the perception of voicing in initial stop consonants was proposed in terms of the discriminability of differences in the temporal order of the component events at stimulus onset. At the time, we argued that these results with nonspeech patterns as well as the earlier data using speech signals with infants, adults, and chinchillas reflect a basic limitation of the ability of the auditory system to process (i.e., identify) temporal-order information in both speech and nonspeech signals (Hirsh, 1959). With regard to the cues to voicing perception in word initial stops as cued by VOT, we suggested that the time of occurrence of an event (i.e., the onset of voicing) must be perceived in relation to the temporal attributes of other events (i.e., the release from stop closure). The fact that these events, as well as others involved in VOT perception, are ordered in time implies that highly distinctive and discriminable changes will be produced at various regions along this temporal continuum. Thus, the discrimination of small temporal differences such as those used here will be poor in some regions of the stimulus continuum whereas the discrimination of discrete attributes across perceptual categories will be excellent. This is exactly what the previous categorical perception experiments demonstrated and reflects fundamental properties of the phonological systems of natural languages. As Stevens and Klatt (1974) observed a number of years ago, the inventory of phonetic features used in natural languages is not a continuous

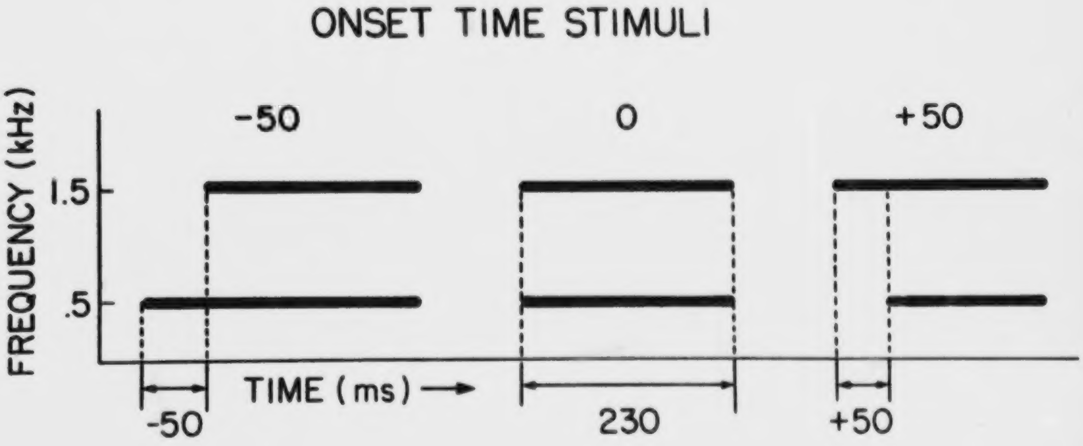

ONSET TIME STIMULI
FREQUENCY (kHz)
1.5
.5
-50
0
+50
TIME (ms) →
-50
230
+50

Figure 1. Schematized displays of nonspeech tone analog stimuli differing in relative onset time of individual components. [From Pisoni, 1977].

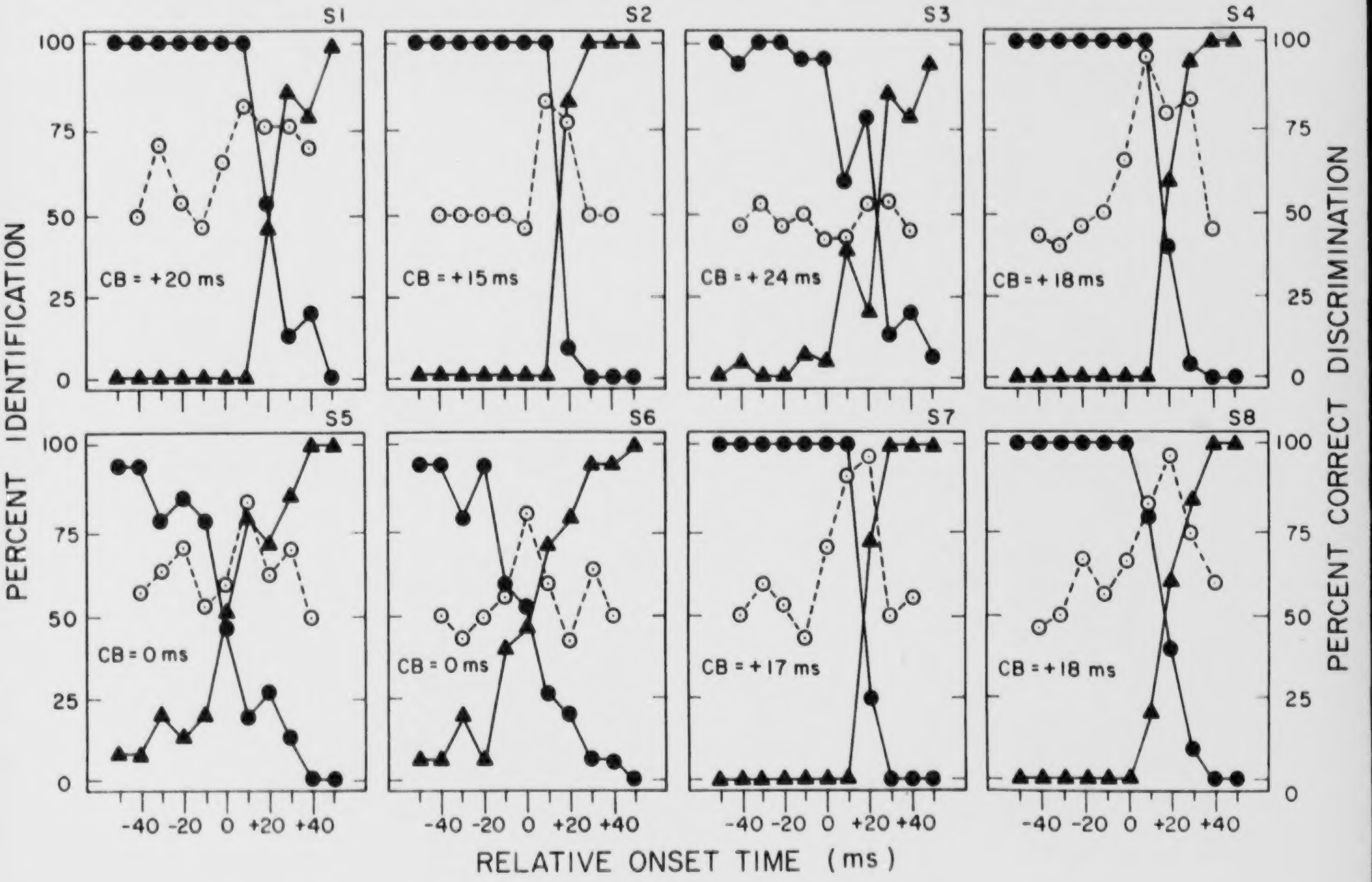


Figure 2. Identification functions (filled circles and triangles) and ABX discrimination functions (open circles) for nonspeech signals differing in tone onset time. [From Pisoni, 1977].

variable but rather consists of the presence or absence of discrete sets of attributes or cues. One of these attributes appears to be related to the perception of simultaneity at stimulus onset.

Perception of the Duration of Rapid Spectrum Changes

For many years speech researchers have been interested in how one phoneme affects the perception of other phonemes in the speech signal. This general phenomena has been called context conditioned variability in speech and it has been a major theoretical issue in the field (see Miller, 1981). Despite the variablity in the physical signal, listeners display a form of perceptual constancy or normalization. Several hypotheses have been proposed over the years to account for this process. One view assumes that listeners track changes in the talker's speaking rate. According to Miller and Liberman (1979), the listener interprets a particular set of acoustic cues in the speech signal, such as the duration of a formant transition for [ba] or [wa], in relation to the talker's speaking rate rather than by reference to some absolute set of context-invariant attributes in the auditory pattern itself. In Miller and Liberman's well-known study on the perception of [ba] and [wa] they found that the labeling boundary for a syllable-initial [b-w] contrast was determined by the overall duration of the syllable containing the target phoneme. Thus, listeners adjusted their decision criteria to compensate for the differences in vowel length that are produced at different speaking rates.

We became interested in these claims concerning the perceptual basis of normalization for speaking rate and carried out a nonspeech control experiment to determine if similar changes also occur when the signals contain rapid spectrum changes but do not sound like speech (Pisoni, Carrell, & Gans, 1983). Examples of the test stimuli are shown in Figure 3.

Insert Figure 3 about here

As in the Miller and Liberman study, we varied stimulus duration of the test pattern and studied the effects of this manipulation on the identification of the duration of a rapid spectrum change at stimulus onset. Subjects were required to identify the onsets of these nonspeech patterns as either "abrupt" or "gradual." The results of our identification study are shown in Figure 4 for both speech and non-speech stimuli. We observed comparable context effects for perception of the duration of rapid spectrum changes as a function of overall duration of the stimulus with both speech and nonspeech signals. Our findings from this nonspeech control study therefore call into question the earlier claims made by Miller and Liberman that context effects such as these are specific to processing speech signals and somehow reflect the listener's normalization for speaking rate.

Insert Figure 4 about here

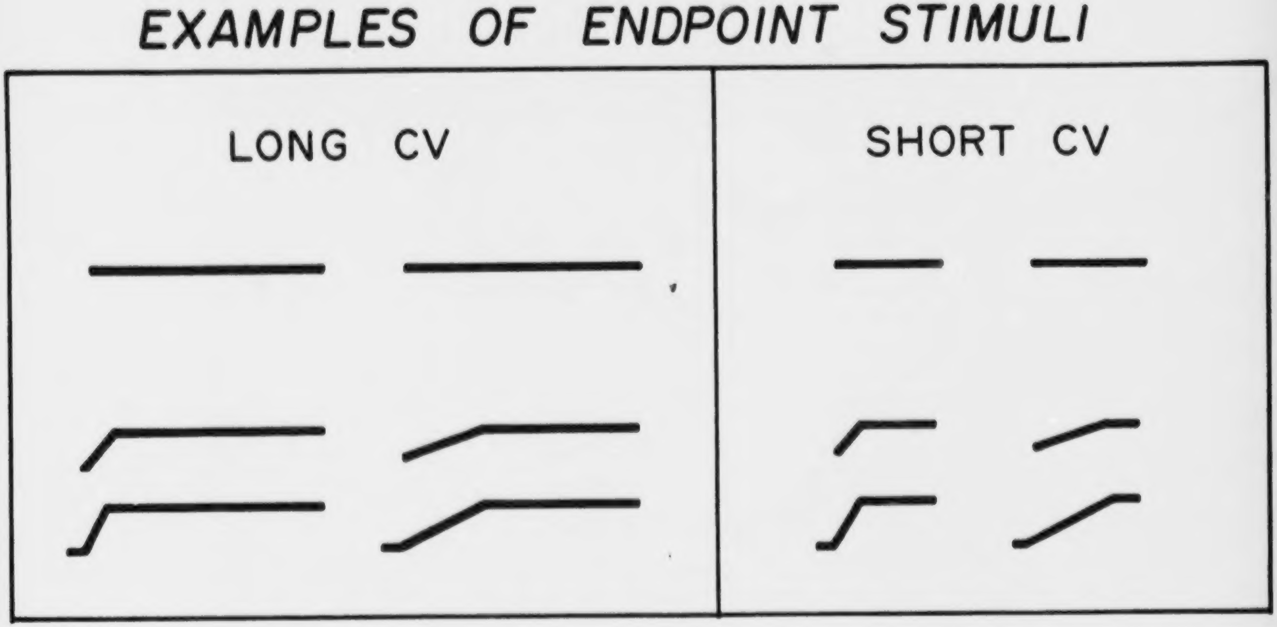

EXAMPLES OF ENDPOINT STIMULI
LONG CV
SHORT CV

Figure 3. Schematized displays of the formant motions of endpoint stimuli corresponding to [ba] and [wa]. Long duration syllables are shown on the left, short syllables are shown on the right. [From Pisoni, Carrell, & Gans, 1983].

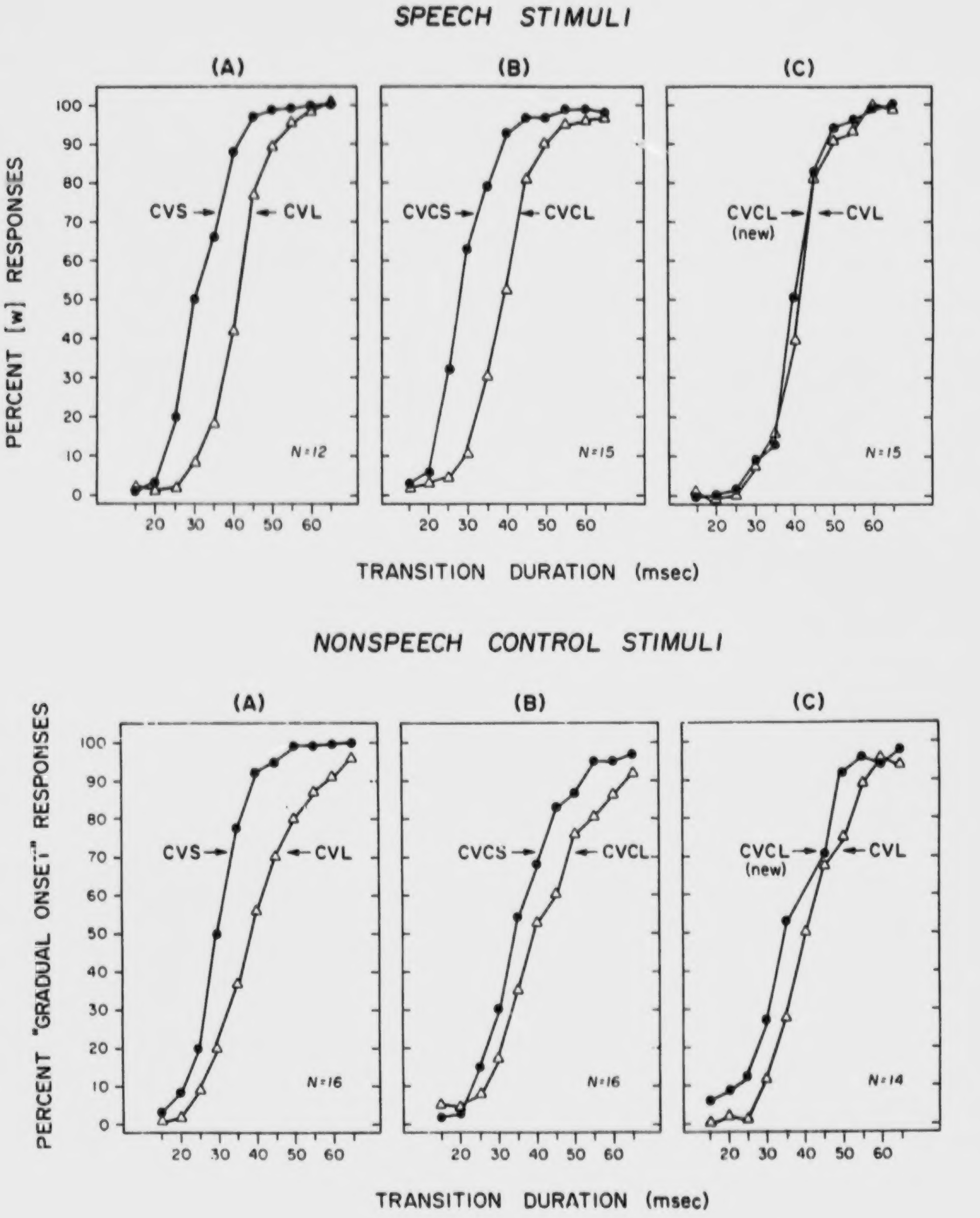


Figure 4. Labeling functions for speech stimuli (top panel) and nonspeech control patterns (bottom panel) that were generated from the displays shown in the previous figure. [From Pisoni, Carrell, & Gans, 1983].

We suggest that context effects such as these may simply reflect general psychophysical principles that influence the perceptual categorizaton and discrimination of all auditory patterns, whether speech or nonspeech. In our experiment, the perceptual categorization of stimulus onsets as either "abrupt" or "gradual" appears to be influenced by later occurring events in the stimulus configuration as observed with speech stimuli. Thus, complex nonspeech signals may also be processed in a "relational" mode, that is, in a manner comparable to that observed in the perception of speech. Our results were particularly striking because we replicated not only the contextual effects reported by Miller and Liberman for syllable duration as displayed in Figure 4 but we also found the same effects as they did when simulated formant transitions were added to the end of the sinusoidal replicas of CV syllables, thus changing the internal structure of the stimulus pattern itself. In short, a relational or nonlinear mode of processing auditory patterns is not limited specifically to the perception of speech signals or to a distinctive phonetic mode of response.

Conclusions

The two sets of findings summarized here taken together with other studies using nonspeech signals suggests that it is possible to offer alternative accounts of specific phenomena observed in speech perception within a somewhat larger context of what is currently known about auditory pattern perception. In the past, it has been very easy to explain a set of findings in speech perception by appealing to the existence and operation of specialized speech processing mechanisms. As we have seen, such global explanatory accounts are no longer satisfactory as we begin to learn more about the psychophysical and perceptual properties of speech and complex nonspeech signals and how the auditory system encodes these types of acoustic patterns. These findings make it clear to us that theoretical accounts of speech perception can no longer be couched in terms of vague descriptions of articulatory mediation via specialized perceptual mechanisms. All of the relevant nonspeech control studies have not been carried out yet but the results of these initial studies are very encouraging that some rapprochement between speech and hearing scientists is possible in the future.

References

Hirsh, I.J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America*, *31*, 759-767.

Miller, J.L. (1981). Effects of speaking rate on segmental distinctions, in *Perspectives on the Study of Speech*, P.D. Eimas & J.L. Miller, (eds.), Lawrence Erlbaum Associates, Hillsdale, NJ.

Miller, J.L., & Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonants and semi-vowels. *Perception & Psychophysics*, *25*, 457-465.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, *61*, 1352-1361.

Pisoni, D.B., Carrell, T.D., & Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, *34*, 314-322.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America*, *24*, 745-749.

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, *68*, 836-842.

Stevens, K.N., & Klatt, D.H. (1974). The role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, *55*, 653-659.

Additional References

Best, C.T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*, 191-211.

Grunke, M.E., & Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, *31*, 210-218.

Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, *1*, 153-167.

Liberman, A.M., Delattre, P.C., Gerstman, L.J., & Cooper, F.S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, *52*, 127-137.

Liberman, A.M., Harris, K.S., Kinney, J.A., & Lane, H.L. (1961). The discrimination of relative onset time of the components of certain speech and non-speech patterns. *Journal of Experimental Psychology*, *61*, 379-388.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., & Halwes, T.G. (1971). Discrimination in speech and non-speech modes. Cognitive Psychology, 2, 131-157.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.

Pisoni, D.B., & Luce, P.A. (1986). Speech perception: Research, theory, and the principal issues, in Pattern Recognition by Humans and Machines, E.C. Schwab & H.C. Nusbaum, eds., Academic Press, New York.

Perceptual Attention in Monitoring Natural and Synthetic Speech*

Howard C. Nusbaum, Steven L. Greenspan, and David B. Pisoni

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

---

*This research was supported in part, by NIH Grant NS-12179, in part by NIH Training Grant T32 NS-07134, and in part by Air Force Contract No. AF-F 33615-83-K-0501 through the Armstrong Aerospace Medical Research Laboratory (AFSC), Wright-Patterson AFB, OH to Indiana University in Bloomington. We thank Michael Stokes and Michael Dedina for their assistance in carrying out the research reported in this paper. Portions of the research included in this report were presented at the 110th meeting of the Acoustical Society of America, Nashville, Tennessee, November, 1985.

Abstract

The role of voice distinctiveness and phonetic discriminability in perception of natural and synthetic speech was investigated. Subjects were instructed to monitor sequences of CV syllables for a specified target syllable in several conditions: (1) targets and distractors produced by the same human talker (N/N); (2) targets produced by a synthetic talker and distractors produced by a human talker (S/N); and (3) targets produced by a synthetic talker and distractors produced by both the same synthetic talker and a natural talker (S/S+N). Results indicate that highly intelligible synthetic targets are detected faster mixed with natural distractors, than are natural targets mixed with natural distractors. However, when subjects are required to discriminate between synthetic targets and synthetic distractors, performance is much worse than for natural targets and natural distractors. The distinctive mechanical sound of synthetic speech only appears to aid perception when there is just a single synthetic message among natural messages. When listeners must discriminate among synthetic messages, performance is significantly worse than when they must discriminate among natural messages.

Perceptual Attention in Monitoring Natural and Synthetic Speech

Under normal circumstances, we are not typically aware of the effort and attention required to recognize spoken language. However, the inability of listeners to fully recognize the linguistic content of two simultaneously presented utterances demonstrates quite clearly that speech perception requires attention (e.g., Bookbinder & Osman, 1979; Moray, 1969; Treisman, 1969). At the same time, it has also been demonstrated that listeners can quickly and accurately detect changes in talker identity or the presentation of a tone within an unattended utterance, even though they are unable to recognize the linguistic content of that utterance (e.g., Cherry, 1953; Lawson, 1966). Clearly then, listeners are able to detect changes in the source characteristics of a signal, even when they do not pay close attention to that signal.

Recently, Simpson and Williams (1980) suggested that the distinctive "sound" of synthetic speech may serve to facilitate its detection among other messages in much the same way. The distinctively different voice quality of synthetic speech compared with natural and coded speech may directly orient the listener to the synthetic message. Simpson and Williams (1980) claimed that "the reason synthesized speech messages may serve as their own alerting signal may be that they possess some perceptual feature that requires only a low level of attention for detection" (p. 328). Therefore, it should be relatively easy to detect a very distinctive or unnatural sounding synthetic utterance in the context of natural speech messages.

However, this hypothesis is based only on the distinctiveness of "voice quality" and does not take into account the acoustic-phonetic structure of the speech. The acoustic cue structure of synthetic speech is impoverished by comparison with the rich and redundant acoustic-phonetic structure of natural speech. As a result, perceptual encoding of synthetic speech may require more attention than natural speech. For example, Luce, Feustel, and Pisoni (1983) reported that perception of highly intelligible synthetic speech generated by MITalk requires more effort and attention than perception of natural speech. This suggests that even though the distinctive quality of a synthetic voice may facilitate detection of synthetic utterances against a background of natural speech, the effort required to recognize the linguistic content of that message may be substantially greater than for natural speech.

Method

To investigate the role of voice quality and segmental intelligibility in message detection and recognition, we used a binaural target monitoring paradigm. Subjects were asked to listen to a sequence of consonant-vowel (CV) syllables and to respond as quickly and accurately as possible whenever a target syllable was heard. At the beginning of each trial, a target syllable (e.g., GA) was presented visually on a CRT display. Following the display of this target syllable, a sequence of 20 CV syllables was presented over headphones, with a 350 msec interstimulus interval. Each sequence of 20 stimuli consisted of six presentations of the target syllable and 14 presentations of different distractor syllables. Target and distractor syllables were drawn from a set of 16 consonants (i.e., /b,d,g,p,t,k,r,l,m,n,j,w,v,z,f,s/) paired with the vowel /a/ produced by a male talker, a female talker, the Paul voice of DECtalk, and the Votrax Type-'N-Talk. These stimuli are shown in the top panel of the first figure.

Insert Figure 1 about here

A different group of subjects participated in each of nine conditions. These conditions are shown in the bottom panel of Figure 1 and result from the combination of three target types with three types of distractors. The targets were either produced by a female talker, DECtalk, or Votrax. For one third of the target conditions, the distractor syllables were produced by a male talker and thus differed in voice from the targets. For another third of the conditions, the targets and distractors were produced by the same voice. Finally, for the remaining conditions, the distractors consisted of a mix of syllables produced by the target voice and syllables produced by the male talker. These conditions are illustrated in Figure 2. In this figure, the underlined syllables represent the target stimuli. Also, syllables that are circled are in a different voice than syllables without circles.

Insert Figure 2 about here

The top panel of Figure 2 shows the condition in which the targets and distractors were presented in different voices. In this condition, subjects can always detect the target using either the voice difference or the identity of the target syllable. Following the logic of Simpson and Williams (1980) target detection should be faster and more accurate as the distinctiveness of the target voice increases. Thus, performance for Votrax targets with natural distractors should be better than performance with the DECtalk targets.

In the middle panel of Figure 2, the target and distractor voices are the same, so subjects can only use syllable identity as a basis for correct target monitoring. In this condition, segmental intelligibility should be the most important factor. Thus, monitoring performance should be best for the natural targets and worst for the Votrax targets, with DECtalk in between.

Finally, at the bottom of the figure, half of the distractors are in the same voice as the target and half are in a different voice. This condition is closest to the actual application for voice response systems, in which there may be several messages presented to an observer. Some of these messages will be background communications that are to be ignored while other messages will be in synthetic speech from the voice response system. The ability of listeners to quickly and accurately monitor for targets in this condition indicates the degree to which the distinctiveness of the target voice facilitates message detection and recognition. Since text-to-speech systems will seldom be used to deliver just a single invariant message to an observer, this condition tests the more realistic case in which there are several messages produced by the synthetic voice and the listener must correctly recognize the target.

Stimulus Set:

BA, DA, GA, PA, TA, KA, MA, NA, RA, LA, WA, YA, FA, SA, VA, ZA

Talkers:

1. Female/Natural
2. Male/Natural
3. DECTalk Paul
4. Votrax Type-'n-Talk

Target/Distractor Conditions:

	TARGETS		
	Female/Natural	**DECTalk**	**Votrax**
Different	Male/Natural	Male/Natural	Male/Natural
Same	Female/Natural	DECtalk	Votrax
Mixed	Female/Natural + Male/Natural	DECtalk + Male/Natural	Votrax + Male/Natural

Figure 1. The top panel shows the talker characteristics and syllables used as stimuli in target monitoring. The bottom panel shows the different testing conditions of the target monitoring paradigm. Three types of targets were presented (Female/Natural, DECTalk, or Votrax) in one of three distractor conditions. The distractors were presented in either the same voice as the targets, a different voice, or a mix of the same voice as the target and a different voice.

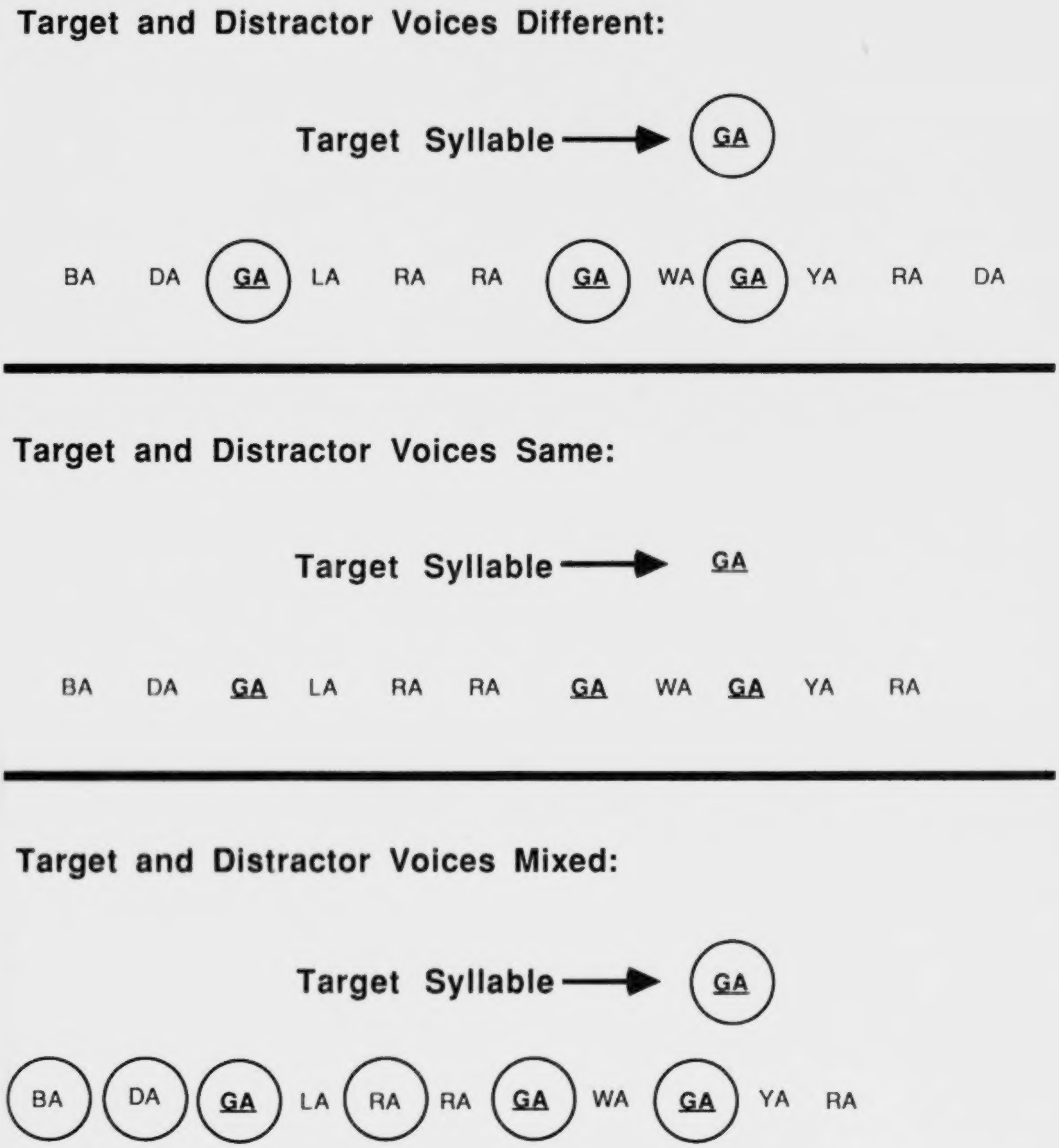

Target and Distractor Voices Different:
Target Syllable GA
BA DA GA LA RA RA GA WA GA YA RA DA
Target and Distractor Voices Same:
Target Syllable GA
BA DA GA LA RA RA GA WA GA YA RA
Target and Distractor Voices Mixed:
Target Syllable GA
BA DA GA LA RA RA GA WA GA YA RA

Figure 2. The three target/distractor conditions in the target monitoring procedure. Targets are shown in boldface and are underlined. Syllables that are circled are presented in one voice and uncircled syllables are presented in a different voice.

Results

Figure 3 shows the percentage of correct target detections in each of the conditions. First, it should be noted that performance is significantly more accurate when the target and distractor voices are different (shown with the open circles) than in the other two conditions. Second, there is no significant difference in performance between the other two conditions. The presence of some distractors in the same voice as the target affects performance as if all distractors are in the target voice. Next, it is important to note that there is a significant decrease in accuracy as a function of target voice, such that natural targets are detected more accurately than Votrax targets, with performance on DECtalk targets in between. Finally, there is a significant interaction indicating that performance on Votrax targets is dramatically impaired when Votrax distractors are presented.

Insert Figure 3 about here

These results indicate that Votrax targets are detected less accurately than natural targets and DECtalk targets in all conditions. Clearly then, the more distinctive sound of the Votrax speech does not provide any advantage in detection, even when all the distractors are natural speech. The response time results further support this conclusion.

Insert Figure 4 about here

Figure 4 shows target monitoring speed for natural, DECtalk, and Votrax targets in the three distractor conditions. Responses are fastest when the target and distractor voices are different. As in the hit rate data, there is little, if any, difference between the Mixed and Same conditions, indicating that the presence of any distractors in the same voice as the targets is sufficient to impair performance. In addition, there is a clear effect of the intelligibility of the target speech such that natural targets are responded to most quickly and Votrax targets are responded to most slowly with performance on DECtalk targets in between. Thus, subjects are fastest and most accurate when detecting natural targets, and slowest and least accurate when detecting Votrax targets regardless of the distractor voice.

Insert Figure 5 about here

The false alarm data shown in Figure 5 support these conclusions, as well. Subjects made fewer false alarms when the target and distractor voices were different than in the other conditions. Also, subjects made fewer false

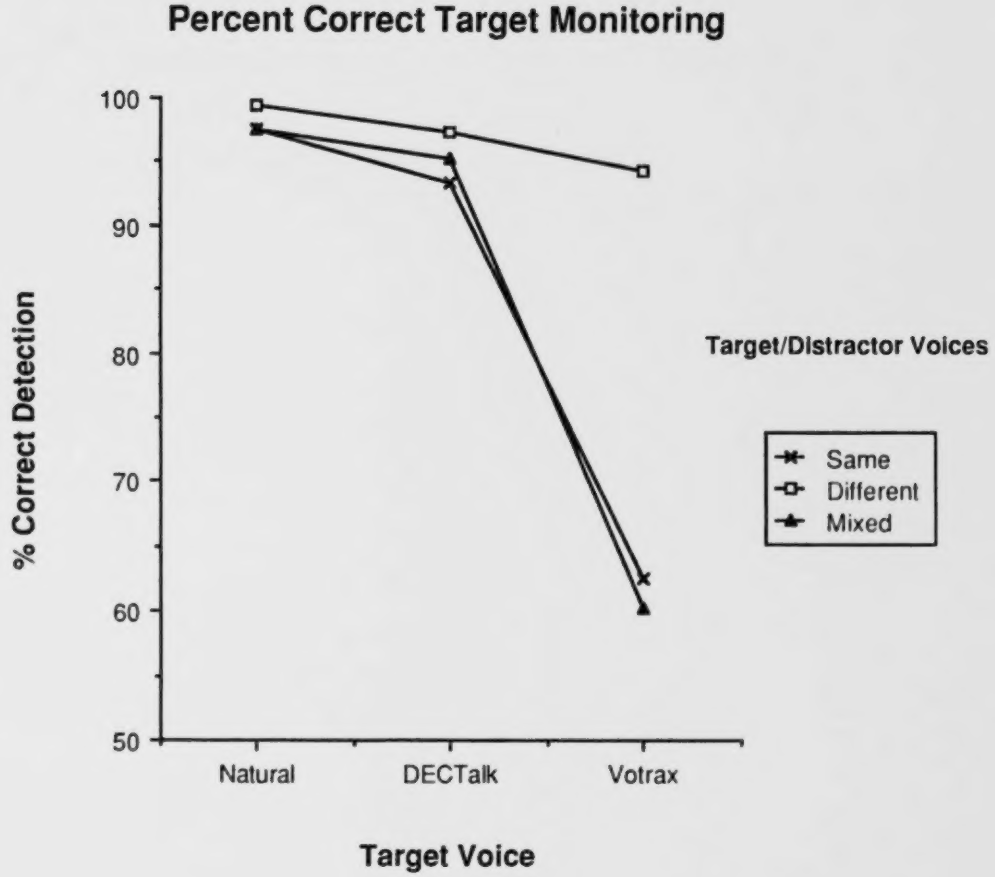


Figure 3. The mean percentage of correct target detection responses averaged across subjects for each of the target and distractor voice conditions.

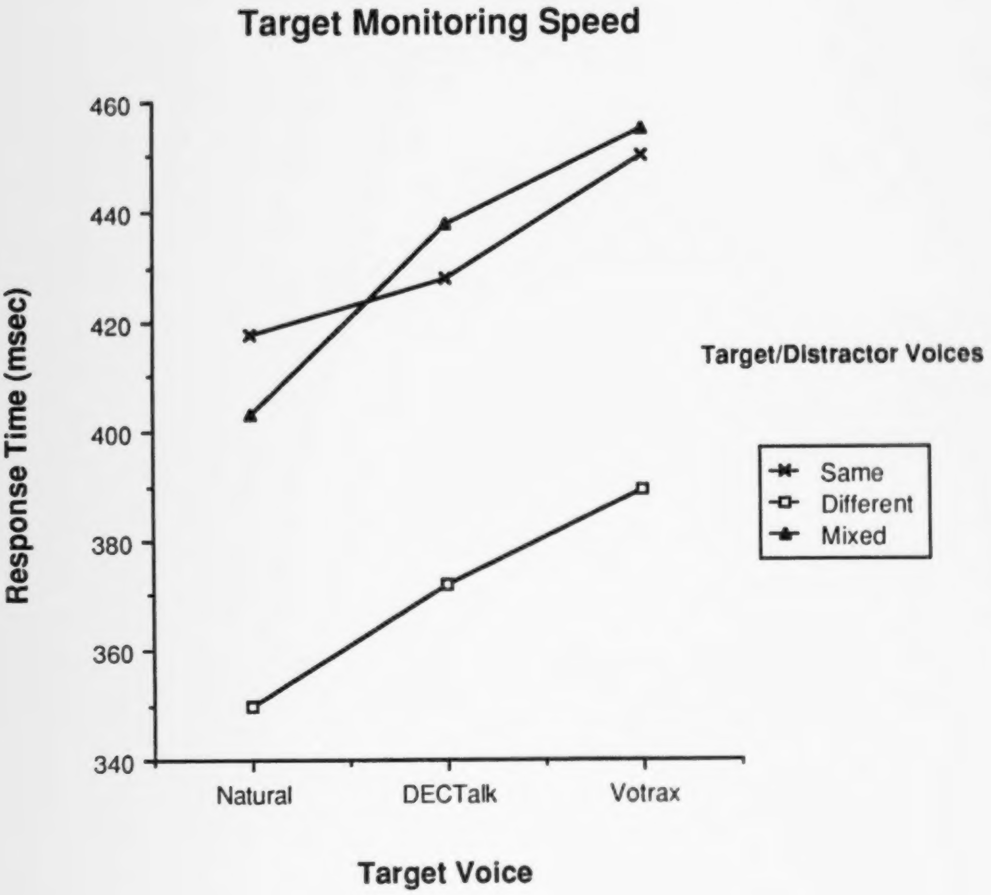


Figure 4. The mean response time in msec. for correct target detection responses in each of the target and distractor voice conditions.

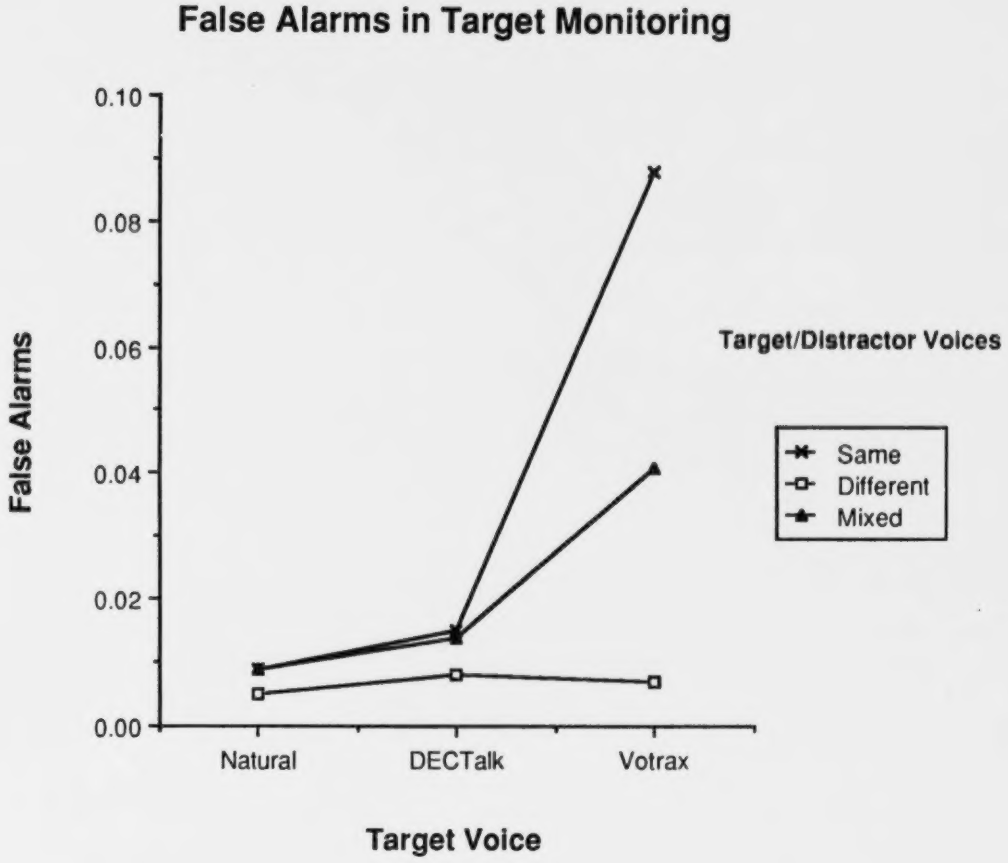


Figure 5. The mean false alarm rate for detection responses to distractor syllables in each of the target and distractor voice conditions.

alarms to natural targets than the synthetic targets. One difference in the pattern of false alarm data compared to the hit rate data can be seen for the Votrax targets. Subjects made half as many false alarms for Votrax targets in the Mixed condition than they did in the condition where all distractors were produced by Votrax. This result occurred simply because there were half as many Votrax distractors in the Mixed condition as in the Same condition, and none of the false alarms were responses to the natural distractors.

Discussion

Taken together these results argue that the effort and attention required for recognizing synthetic speech is not reduced when the synthetic speech must be detected against a background of speech produced by other talkers. Rather, recognition of synthetic speech seems to be impaired, by comparison with natural speech, regardless of the conditions under which it is presented. Similarly, Pierce and Remington (1984) reported that natural speech flight instructions were more intelligible than synthetic speech when presented against a background of air traffic control communications. Moreover, this result was found across several signal-to-noise ratios and after several days of repeated exposure.

The overall pattern of our results indicates that the distinctiveness of a voice is less critical to target detection than the segmental intelligibility of the speech. This argues against the hypothesis that synthetic speech possesses acoustic features that require only a low level of attention for detection. Instead, it is apparent that synthetic speech is harder to detect in the context of natural speech than speech produced by a different natural talker. There is no special perceptual feature in synthetic speech that reduces the effort required for detection. Indeed, it appears that highly distinctive synthetic speech is more difficult to detect than more natural sounding synthetic speech.

One surprising finding was that the speed and accuracy of target monitoring depended on the intelligibility of the speech, even when subjects could have detected targets based on the difference in target and distractor voices alone. This result suggests that subjects processed the phonetic content of the target stimuli despite the fact that it was not necessary for performing the task. Subjects may have been unable to ignore the segmental information encoded in these targets.

In conclusion then, the results of the present study support the claim that recognition of synthetic speech requires more effort and attention than natural speech. Moreover, there is no advantage in detecting synthetic speech against a background of natural speech indicating that there is no special alerting property in synthetic speech. Thus, there appears to be no perceptual advantage for using low-quality synthetic speech in voice-response systems in high-information-load applications and high-noise environments. However, there is still much research that is needed to determine how noise and cognitive load interact in perception of synthetic speech and how training and experience affect the effort required to recognize utterances generated by a text-to-speech system.

References

Bookbinder, J., & Osman, E. (1979). Attentional strategies in dichotic listening. *Memory & Cognition*, *7*, 511-520.

Cherry, E. E. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America*, *25*, 975-979.

Lawson, E. (1966). Decisions concerning the rejected channel. *Quarterly Journal of Experimental Psychology*, *18*, 260-265.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, *25*, 17-32.

Moray, N. (1969). *Listening and attention*. Baltimore: Penguin.

Pierce, L., & Remington, R. (1984). Comprehension of synthetic and natural speech in the presence of competing human speech. *Proceedings of the Voice Data Entry Systems Applications Conference*. Palo Alto, CA: AVIOS.

Simpson, C. A., & Williams, D. H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. *Human Factors*, *22*, 319-320.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, *76*, 282-299.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Intelligibility of Phoneme Specific Sentences Using Three Text-to-Speech Systems and a Natural Speech Control*

John S. Logan and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47401

---

*The research reported in this paper was supported, in part, by NIH Research Grant NS-12179 and, in part, by a contract with the Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH. We thank Paul Luce for providing the natural speech used in this experiment.

Abstract

The performance of three text-to-speech systems and a natural speech control was assessed using the Phoneme Specific Sentences (PSS) developed by Huggins & Nickerson (1985). Use of the PSS stimuli enabled us to compare the intelligibility of different classes of phonemes produced by the four voices in sentence contexts. Subjects were asked to transcribe the sentences they heard as accurately as possible. Transcriptions were scored according to an exact match criterion. Analyses revealed differences among the overall intelligibility of the voices. As expected, natural speech was significantly more intelligible than any of the three synthetic voices. However, examination of the errors across the different phoneme categories revealed different response patterns among the voices. The results demonstrate that the PSS stimuli provide valuable information about the quality of phoneme synthesis in sentence contexts. The present results also replicate the findings of earlier studies carried out in our laboratory showing qualitative differences among the patterns of perceptual confusions that occur when listening to different types of synthetic speech and those that occur when listening to natural speech. Taken together, our results provide further evidence for the claim that synthetic speech produced by rule displays an impoverished acoustic-phonetic structure compared to natural speech. Our results also show important differences among several synthetic voices suggesting that synthetic speech, like natural speech, should not be viewed collectively as a generic entity.

Intelligibility of Phoneme Specific Sentences Using Three Text-to-Speech Systems and a Natural Speech Control

The perceptual evaluation of devices that process (i.e., encode and transmit) speech signals has been the object of considerable effort ever since Fletcher & Munson (1929) determined that the perceived quality of voice transmission over telephone circuits could not be determined solely through electronic measurement techniques. Over the years, a relatively large number of perceptual tests have been developed that provide useful information on several characteristics of the speech processed by devices ranging from radios to vocoders and text-to-speech systems. With respect to the latter two devices, most of these tests have examined segmental intelligibility since, until recently, this factor had been assumed to be most critical to the perception and understanding of speech. Although this emphasis on the assessment of segmental intelligibility was well placed, other factors, such as preference and naturalness, are also likely to play an important role in determining the acceptability of speech processing devices to listeners.

Recently, Huggins & Nickerson (1985) described the development of a set of Phoneme Specific Sentences (PSS) that they had found useful for evaluating the preferences subjects displayed for different types of vocoded speech. Each sentence contained a large proportion of words containing a set of phonemes corresponding to a specific class, such as voiced stops, nasals, liquids or glides. They used a small set of these PSS stimuli to examine the preferences subjects had for speech processed by several different vocoders. Huggins & Nickerson found the PSS stimuli were useful materials for reliably differentiating among the preferences that subjects had for different vocoders. Although Huggins & Nickerson (1985) used only a small set of these stimuli in their preference experiment, they included nearly one hundred sentences in their report that were constructed using similar principles as those sentences they actually used.

The usefulness of these stimulus materials to our ongoing studies on the perception of synthetic speech became apparent to us after carrying out an experiment in which we examined the preferences subjects had for different synthesized voices obtained from several text-to-speech systems (see Logan & Pisoni, 1986). In our first experiment, we used a small set of Harvard sentences (Egan, 1946) as stimuli and found that intelligibility appeared to be a good predictor of the preferences subjects display for one synthetic voice over another. In a second study, we wanted to see if this effect was reliable using different stimulus materials. To this end, we carried out another experiment in which we used the PSS stimui developed by Huggins & Nickerson (1985). The reason we chose these particular sentences was because the results of the Huggins & Nickerson experiment indicated these sentences might be useful stimuli in evaluating preferences for different types of vocoded speech. Furthermore, the information provided by the different phoneme categories and how they were related to preference was also of interest. The results of our second preference experiment replicated the findings obtained in the first experiment. Preference was positively correlated with the pattern of differences in intelligibility.

The measure of intelligibility that we used in our preference study was the Modified Rhyme Test (MRT) developed by House, Williams, Hecker, & Kryter (1965). This test was designed to assess the intelligibility of individual isolated monosyllabic words. In light of the obvious differences in performance between isolated words and words in sentences, we decided to collect intelligibility data directly from the PSS stimuli generated by the three synthetic voices used in the earlier preference experiment. Furthermore, to provide a benchmark against which to judge the intelligibility of the sentences tested with synthesized speech, we also used a natural speech control.

In short, in this study, we examined transcription performance for the PSS output from three text-to-speech systems (DECtalk, Prose, and Infovox) and natural speech obtained from a native speaker of English. We were interested in the variations among the four voices in the transcription accuracy across the 18 individual phonetic categories used in the PSS stimuli. Another factor that we examined was the gross structural characteristics of the sentences and how they may have affected intelligibility. The structural component that we chose to look at was the number of words in a sentence which can be taken as a rough estimate of syntactic complexity.

Method

Subjects. For the three conditions of the experiment in which output from text-to-speech systems was used, thirty subjects (ten for each condition) were recruited from a paid subject pool maintained by the Speech Research Laboratory. These subjects were paid $3.50 for their participation. For the condition in which natural speech was used, ten subjects were recruited from a volunteer subject pool maintained by the Department of Psychology at Indiana University. The subjects in the natural speech condition received course credit for their participation. All subjects used in these tests were native speakers of English and reported no history of a speech or hearing disorder at the time of testing. All subjects were drawn from the same general population of undergraduate students at Indiana University.

Stimuli. A subset of the Phoneme Specific Sentences developed by Huggins & Nickerson (1985) were used as stimuli in the present experiment. Ninety-two sentences were selected for use in this study. The sentences are given in Appendix A. The synthesized stimuli used in the present experiment were obtained from the digitized waveform files used in our earlier preference experiment. These waveform files were produced in the following way. The 92 PSS stimuli were generated using the default voices for three text-to-speech systems, DECtalk V3, Prose 2000 V3, and Infovox SA101. The PSS output from each text-to-speech system was recorded on audiotape, low pass filtered at 4.8 kHz, and then digitized on a PDP-11/34 computer at a 10kHz sampling rate using an A/D converter with 12 bit resolution. Individual waveform files for each sentence were created using WAVES, a waveform editing program (see Luce & Carrell, 1981). After digitization, each sentence was processed by a level adjustment program (Bernacki, 1981) in order to ensure an approximately 50 dB RMS level across each sentence.

The natural speech stimuli were produced by a male talker (PAL) who read the sentences at a moderate speaking tempo from a randomized list. The sentences were recorded in an IAC sound-attenuated booth on a Crown 800 Series tape recorder. The recordings were digitized and segmented into individual waveform files for each sentence. The natural speech was also processed by a level adjustment program in order to obtain an approximately 50 dB RMS level across each sentence.

Procedure. Subjects were tested in individual listening booths in groups of four to six. The stimuli were presented over matched and calibrated TDH-39 headphones at a level of approximately 80dB SPL. The amplitude of the test stimuli was measured in relation to a calibration signal, a 10s vowel /a/, produced by DECtalk. Measurements were carried out using a VTVM. Individual audiotapes containing the stimuli produced by each voice were reproduced using an Ampex AG500 tape recorder. White noise of approximately 50dB amplitude was mixed with the speech signal to mask background and tape noise. The tape recorder and associated equipment were located in an adjacent room and were remotely controlled by the experimenter who remained with the subjects in the testing room.

Each group of subjects was told that they would be listening to English sentences and that their task was to write down each sentence they heard. Subjects were also told that if they were unsure about what they heard, they should write down whatever they thought they heard, even if they had to guess. Subjects recorded their responses in specially prepared response booklets. After each sentence was presented, the experimenter stopped the tape recorder so that subjects could write the sentence down. After all subjects had indicated they had recorded their response, the next sentence was presented. This procedure was continued until all the sentences had been presented. The experiment required approximately 45 minutes to complete.

The transcription responses were scored according to the following criteria: Spelling errors were scored as correct if there was a phonemic match between the response and the actual stimulus presented. Otherwise, sentences were scored as correct only if there was an exact match between the response and the intended stimulus. Omissions, transpositions, and additions were all scored as errors. In other words, the entire sentence was required to be completely correct in order for it to be scored as correct.

Results

Overall Percent Error

The overall proportion of error responses for each voice is shown in Figure 1. The lowest error rate as measured by transcription accuracy was obtained for natural speech, followed by DECtalk, then Prose, and finally, Infovox. The differences in error rate between the natural speech control and the three synthetic voices are very apparent from this figure. In contrast, the differences between DECtalk and Prose appear very small, while the differences between Infovox and the other two synthetic voices appears to be quite large.

Insert Figure 1 about here

In order to confirm the trends observed in Figure 1, an analysis of variance was carried out. The between subjects factor was voice (four voices) while the within subjects factor was phonemic category (18 categories). Significant main effects for voice [$F(3, 36)=119.73$, $p<0.0001$] and phonemic category [$F(17, 612)=26.7$, $p<0.0001$] were obtained. In addition, a significant interaction between voice and phonemic category was also obtained [$F(51, 612)=4.51$, $p<0.0001$]. The main effect of voice was further examined to determine which voices were significantly different from each other. Newman-Keuls tests comparing the mean percent error indicated that the performance obtained with the natural speech was significantly better ($p<0.05$) than any of the synthetic voices. Furthermore, the performance of both Prose and DECtalk was significantly different from that of Infovox. No significant

324

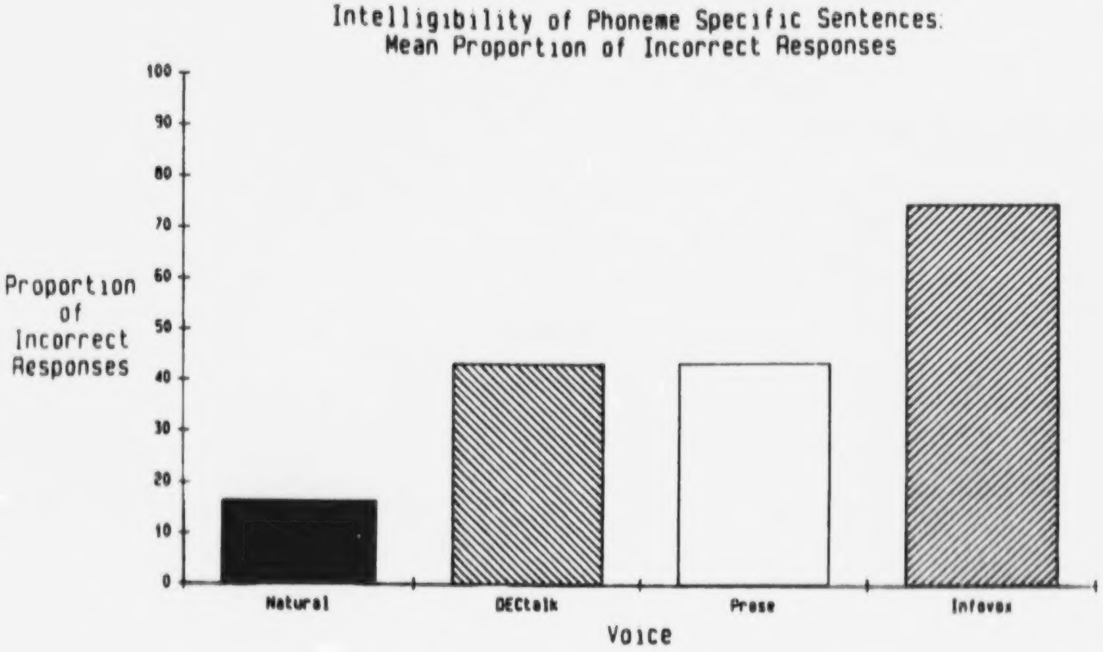


Figure 1. Overall proportion of errors for each voice.

differences were found, however, in the transcription performance between DECtalk and Prose. Thus, the trends observed in Figure 1 were confirmed.

It is of some interest to compare the results obtained using the PSS stimuli described above with results obtained in our laboratory using other types of stimulus materials. In earlier studies examining the intelligibility of synthetic speech (e.g., Greene, Logan, & Pisoni, 1986; Logan, Pisoni, & Greene, 1985; Greene, Manous, & Pisoni, 1984), we found that the overall error rates for synthetic speech were consistently higher than those obtained with natural speech. This difference in intelligibility between natural and synthetic speech was observed consistently across several different types of stimulus materials, including CV syllables, isolated words (MRT), and words in sentences (Harvard and Haskins sentences). Not surprisingly, this effect was replicated in the present experiment; that is, the natural voice was more intelligible than any of the synthetic voices. However, some of the relative differences in performance that existed among the different synthetic voices obtained using measures such as the MRT changed when the PSS stimuli were used. Specifically, a statistically reliable difference in overall error rate was found between DECtalk and Prose when the MRT was used as the measure of intelligibility (see Logan, Pisoni, & Greene, 1985). In contrast, no reliable difference was observed in error rates between DECtalk and Prose using the PSS stimuli in this study. The differences in results will be considered briefly below.

Error Analysis by Phoneme Class

The data were also tabulated and analyzed according to the different phonetic categories used to construct the PSS stimuli. These data are shown in Table 1. The percentage of error responses is shown for each phonetic category and for each voice. The significant interaction obtained in the analysis of variance described above suggested differences in performance among the voices as a function of phonetic category. Simply put, different voices exhibited different patterns of errors.

Insert Table 1 about here

In order to obtain a more detailed assessment of the relationship among the error patterns of the different voices, Pearson product moment correlation coefficients were calculated for the six possible pair-wise combinations. These correlations are shown in Table 2. The highest correlation was obtained between DECtalk and Prose, followed by the correlation between DECtalk and Infovox. The lowest correlation was observed between Prose and natural speech, and the next highest was between Infovox and natural speech. All of the correlations were significantly different from zero ($p<0.05$) except the correlation between Prose and natural speech. No significant differences between the correlations coefficients were obtained.

Insert Table 2 about here

Table 1

Proportion of Error Transcription Responses for Phoneme Specific Sentence Categories

PSS Categories	Proportion of Error Responses (%)			
	DECtalk	Prose	Infovox	Natural
1) all fricatives.............	60.2	60.00	93.33	13.33
2) all stops & affricates......	30.00	56.67	80.02	0.00
3) all consonant phonemes......	73.33	85.00	98.33	44.00
4) glides except l & vowels....	50.00	60.00	75.00	1.67
5) glides......................	35.00	5.00	70.00	45.00
6) glides & nasals.............	38.00	23.99	68.67	7.01
7) all labials.................	27.50	42.50	75.00	3.17
8) nasals......................	12.87	20.02	67.13	2.86
9) nasals + l..................	20.00	5.00	70.00	6.43
10) all tongue tip.............	74.26	57.13	92.85	34.29
11) all unvoiced consonants....	72.00	42.00	92.00	16.00
12) unvoiced fricatives........	45.00	45.00	80.00	15.00
13) unvoiced stops.............	26.00	42.00	42.00	2.00
14) unvoiced stops & affricate.	44.00	47.00	47.00	6.00
15) voiced fricatives..........	80.00	77.50	100.00	37.50
16) all voiced consonants......	24.00	40.00	74.00	24.00
17) voiced stops...............	23.75	28.75	61.25	7.75
18) voiced stops & affricate...	50.00	50.00	65.00	30.00
Overall	43.65	43.76	75.09	16.41

Table 2

Correlations Between Voices for Transcription Accuracy in PSS Phonetic Categories

	DECtalk	Prose	Infovox	Natural
DECtalk	1.00	0.722	0.712	0.590
Prose	-	1.00	0.535	0.273
Infovox	-	-	1.00	0.506
Natural	-	-	-	1.00

We also carried out an analysis to assess the effects of sentence length on transcription accuracy. Because there was a great deal of variability in the number of words used in the 92 sentences, which ranged in length from four to fifteen words, we wanted to determine the extent that sentence length affected transcription accuracy. For each voice, Pearson product moment correlations were calculated to assess the relationship between sentence length and transcription accuracy. The correlation coefficients were 0.250 for DECtalk, 0.256 for Prose, 0.328 for Infovox, and 0.368 for natural speech. All of the correlations were significantly different from zero ($p<0.05$). Thus, longer sentences tended to be transcribed less accurately than shorter sentences. This result was observed for natural speech as well as synthetic speech.

Discussion

The results of the present study suggest two main conclusions. First, the PSS stimuli are more difficult than other stimulus materials we have used in our previous studies on the perception of synthetic speech, such as the MRT vocabulary or the Harvard and Haskins sentences. This conclusion is supported by the higher error rates observed for all voices using the PSS stimuli. Obviously, the differences between stimuli comprised of individual isolated words, such as those used in the MRT, and stimuli comprised of highly complex and variable sentences, such as those used in the PSS, are not surprising and are suggestive of the diagnostic utility of these materials in speech perception studies. With sentences such as these, it is possible to identify sources of error and associate them with specific classes of phonemes that may be synthesized poorly or even incorrectly in a particular context.

One example of the diagnostic usefulness of the PSS materials emerges from the finding that natural speech was more intelligible than synthetic speech. Since this result was obtained for even DECtalk, the most intelligible of the synthetic voices, even small improvements in synthesis quality may be revealed by comparing the performance of synthetic speech and natural speech using the PSS. The generality of this explanation of how small differences in intelligibility may be exploited to infer improvements in synthesis, however, must be qualified by noting that the differences in transcription performance between DECtalk and Prose became negligible when we consider only the gross overall measure of error responses. Despite the fact that the overall error rates for DECtalk and Prose were comparable, the results of the present study demonstrated quite clearly that the source of errors for DECtalk and Prose were very different.

The high error rates obtained using the PSS stimuli also points out the somewhat arbitrary effects of different stimulus materials on intelligibility scores. The effect of different stimulus materials may be further illustrated by considering the influence of sentence length on error rates. In the present study, we found that sentence length was positively correlated with error rate. Other structural factors, such as word frequency, familiarity, syntactic structure, and semantic coherence also may have contributed to the low overall performance obtained using the PSS stimuli. Some of these structural differences were necessary to satisfy the constraints imposed by creating sentences loaded with specific phonemes without the sentences becoming completely anomalous.

The second major conclusion that may be drawn from the results of the present investigation is that large differences exist among the types of errors found in natural and synthetic speech using these specially constructed materials. Not only were large differences in the error patterns observed between the natural and synthetic speech, but differences were also observed among the error patterns for different synthetic voices. For example, even though the speech produced by DECtalk is highly intelligible, it is still quite different from natural speech in many important ways, as shown by an examination of the error patterns and the correlations with natural speech. A similar conclusion was suggested in earlier work done in our laboratory. Several studies showed that the patterns of perceptual confusions were different for natural and synthetic speech, suggesting that synthetic speech was not simply the same as natural speech degraded by noise (see Nusbaum, Dedina, & Pisoni, 1984; Yuchtman, Nusbaum, & Pisoni, 1985). Indeed, the results implied that synthetic speech had an impoverished acoustic-phonetic cue structure. Some cues to phonemic contrasts were present but others were distorted or missing in the phonetic implementation rules used in synthesis.

Thus, one of the major difficulties with synthetic speech produced by rule lies in the manner in which the phonetic information is actually converted into an acoustic waveform, the phonetic implementation rules. In addition, factors such as prosody and naturalness undoubtedly play an important role in determining the intelligibility of a particular sequence of synthesized speech. However, the present error analyses strongly point to segmental intelligibility as a source of the differences in transcription performance. Until all these factors are better understood, synthesized speech will continue to remain less intelligible than natural speech. We believe it is important to emphasize here that synthetic speech, like natural speech, is not a homogenous entity and that it may be misleading to talk about synthetic speech as if it were simply a generic form of speech produced by rule, as some researchers have implied. The present tests establish quite firmly, in our view, that important differences exist among different kinds of synthetic speech produced by rule. And, of course, consistent differences between the perception of natural speech and various kinds of synthetic speech continue to be observed using a wide variety of stimulus materials. Only by understanding the nature of these differences and their locus will we be able to improve the quality and intelligibility of synthetic speech produced by rule, so that it eventually sounds indistinguishable from natural speech generated by real talkers.

References

Bernacki, R. (1981). WAVMOD: A program to modify digital waveforms. In *Research on Speech Perception. Progress Report No. 7*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Egan, J. P. (1947). Articulation testing methods. *Laryngoscope*, *58*, 955-991.

Fletcher, H, & Steinberg, J. (1929). Articulation testing methods. *Bell Systems Technical Journal*, *8*, 806-854.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers*, *18*, 100-107.

Greene, B. G., Manous, L. M., & Pisoni, D. B. (1984). Perceptual evaluation of DECtalk: A final report on version 1.8. In *Research on Speech Perception. Progress Report No. 10*, Bloomington, IN: Speech Research Laboratory, Indiana University.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, *37*, 158-166.

Huggins, A. W., & Nickerson, R. S. (1985). Speech quality evaluation using phoneme-specific sentences. *Journal of the Acoustical Society of America*, *77*, 1896-1906.

Logan, J. S., & Pisoni, D. B. (1986). Preference judgements comparing different synthetic voices. In *Research on Speech Perception. Progress Report No. 12*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Logan, J. S., Pisoni, D. B., & Greene, B. G. (1985). Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. In *Research on Speech Perception. Progress Report No. 11*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Carrell, T. D. (1981). Creating and editing waveforms using WAVES. In *Research on Speech Perception. Progress Report No. 7*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Nusbaum, H. C., Dedina, M. J., & Pisoni, D. B. (1984). Perceptual confusions of consonants in natural and synthetic speech. In *Research on Speech Perception. Progress Report No. 10*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Yuchtman, M., Nusbaum, H. C., & Pisoni, D. B. (1985). Consonant confusions and perceptual spaces for natural and synthetic speech. Paper presented at 110th meeting of the Acoustical Society of America, Nashville, November, 1985.

Appendix A

Phoneme Specific Sentences

(from Huggins & Nickerson, 1985)

Fricatives -
His vicious father had seizures.
Whose shaver has three fuses?
Three of the chefs saw the thieves.

Stops and affricates -
Which tea party did Judge Baker go to?
We'd better buy a bigger dog.
Georgie had to chew tobacco.

Consonants -
If the treasure vans got so much publicity we think you should hide your share.
The voyagers have ground the crankshaft with unimpeachable precision.
The old-fashioned jacket was giving you both so much humourous pleasure.
The average disillusioned gambler thinks he wishes for a cheap yacht.
Nothing could be further from reality than his illusion of chasing your gorgeous sheep away.
She thinks even the pale rouge you bought was much too gaudy for her age.

Glides except l -
Why were you away a year Roy?
Why were you weary?

Glides -
Our lawyer will allow your rule.
Our rule will allow you a lawyer.
We really will allow you a ruler.

Glides and nasals -
You were wrong all along.
I know you're all alone.
When will our yellow lion roar?
An alarm rang a warning in only one room.
A lawyer may well allow a new ruling.
I'm learning my new role.
I'll remain in my narrow room.
Anyone may rely on a mailman.
I'm wearing my maroon ring.
We'll allow you a new loan.
I'll lie in an alarming manner.
Why lie when you know I'm your lawyer?
A normal animal will run away.
Mail me an aluminum railing.
I'll willingly marry Marilyn.

Appendix A (continued)

Phoneme Specific Sentences

Labials -
Pay my wife by five.
Weave me a web above a poppy.
Move off my pew baby!
Weep for my baby puppy.

Nasals
Nanny may know my meaning.
I'm naming one man among many.
No one knows my name.
I know many a mean man.
I know no minimum.
Many young men owe money.
When may we know your name?

Nasals plus l -
I'm well known among men.
Nine men moaning all morning.

Tongue tip -
The judge's short decision really touched the youth.
Each decision shows the jury she lies through her yellow teeth.
Such a rash allusion to dosage teases the youth.
Seth yawns at each rash allusion to the dosage.
The designers really earned the judge's derision this year.
Each allusion to Daisy's agility lessens her attention.
Each decision shows that he lies through his yellow stained teeth.
John drowned his sorrows in gin and orange juice.

Unvoiced consonants -
She swiftly passed a health check.
He steps off a path to cash a check.
I hope she chased her fox to earth.
A thick-set officer pitched out her hash.
He checked through fifty ships.

Unvoiced fricatives -
A thief saw a fish.
I saw three fish.
Three chefs face a thief.

Appendix A (continued)

Phoneme Specific Sentences

Unvoiced stops -
Take a copy to Pete.
Pat talked to Kitty.
Quite a cute act.
Peter took out a potato.
Kate typed a paper.

Unvoiced stops and affricates -
Chip took a picture.
A teacher patched it up.
Chat quietly to teacher.
Quite quiet at church.
Catch a paper cup.
Actuate a paper copier.
A teacher taped up a packet.
Capture a cute puppy.
A teacher typed up a paper.
Katie tacked up a cute picture.

Voiced fricatives -
They use our azure vials.
There's our azure vial.
There's usually a valve.
Those waves veer over.

Voiced consonants -
Does John believe you were measuring the gun?
Your brother's vision was gradually dimming.
The regular division was led by a young major.
I gather you will be abandoning the major revisions?
The young major's evasions were growing bolder.

Voiced stops -
Bobby did a good deed.
I begged Dad to buy a dog.
Did Bobby do a good deed?
Buy Dad a bad egg.
Dad would buy a big dog.
Why did Gay buy a bad egg?
Do you abide buy your bid?
Grab a doggie bag.
A greedy boy died.

Voiced stops and affricates -
Did George do a good job?
Greg adjudged Bobby dead.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

PRONOUNCE: A Program for Pronunciation by Analogy*

Michael J. Dedina and Howard C. Nusbaum

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*The work was supported, in part, by NIH Research Grant NS-12179 to Indiana University in Bloomington. This paper was presented at the Fourteenth Annual ACM Computer Science Conference, Cincinnati, OH, February, 1986.

Abstract

Humans can pronounce novel orthographically-regular text strings such as pseudowords, like "peemers", or words they have never seen before. How do they do this? Two hypotheses have been proposed to account for this ability. According to one view, pronunciation-by-rule, pseudowords are pronounced by a ruled-based phonological process in which the pronunciation for a pseudoword is generated from its spelling by the use of a complex set of spelling-to-sound rules (e.g., Forster & Chambers, 1973). According to the alternative view, pronunciation-by-analogy, pseudowords are pronounced by analogy to known words which are similar in spelling (e.g., Glushko, 1981). Although the pronunciation-by-analogy approach is psychologically plausible, it is not clear that it is computationally feasible. Pronunciation-by-analogy depends on the degree to which orthographic consistency in the spelling patterns of words is related to phonotactic consistency in pronouncing those words. To investigate this theoretical issue, we developed a computer program called PRONOUNCE that automatically generates a set of rank-ordered pronunciations, in the form of a sequence of phonetic segments, using pronunciation-by-analogy with a lexicon of approximately 20,000 words based on Webster's Pocket Dictionary. PRONOUNCE examines every word in the lexicon and builds a pronunciation lattice structure using the phonetic representations of the words that match the input string. In this pronunciation lattice, each node represents a possible phoneme to be used at a particular position in the pronunciation, and each path through the lattice represents a possible pronunciation. At this time, PRONOUNCE performs reasonably well, generally producing pronunciations that agree with those given by native speakers of English. PRONOUNCE was tested on a set of 70 short pseudowords and was found to disagree with human subjects on only 9% of the pseudowords. These results suggest that pronunciation-by-analogy is indeed computationally feasible. Furthermore, the limited success of PRONOUNCE suggests a new approach to spelling-to-sound conversion for text-to-speech conversion systems.

PRONOUNCE: A Program for Pronunciation by Analogy

I. INTRODUCTION

In reading aloud, most people have very little trouble pronouncing novel words and pseudowords. A pseudoword is a string of letters that does not spell a real word, but still conforms to the spelling patterns of English.

Insert Figure 1 about here

Figure 1 shows several example pseudowords that most native English speakers can easily pronounce. If text-to-speech systems could pronounce pseudowords as accurately as most humans do, synthetic speech generated automatically by computer would be more intelligible. A text-to-speech system is a speech synthesizer that converts unrestricted English text in ASCII format into speech. When pronouncing known words, a text-to-speech system will often reference a pre-compiled dictionary containing pronunciations for a number of words. However, unknown spelling patterns that are not found in the dictionary must be pronounced using a relatively large and complex set of spelling-to-sound rules. In the more sophisticated text-to-speech systems currently available, such as DECtalk, Prose, and Infovox, this approach works quite well. However, the development of a set of spelling-to-sound rules for a particular dialect or language requires a great deal of time and the services of an expert linguist. The linguist must analyze the pronunciations of a large corpus of words and then apply explicit linguistic knowledge and intuition to formalize the relationships between spelling and sound patterns for this data. Thus, the development of spelling-to-sound rules by linguistic experience and intuition is a complex process with many opportunies for error that may be compounded by the size of the rule set and side effects among rules. One way to reduce the chances of errorful rules and improve the pronunciations of text-to-speech systems is to somehow automate the role played by the expert linguist. The success of this approach depends, in part, on gaining a better understanding of the way humans pronounce words, especially novel words they have never seen or heard before.

Psychologists have generally assumed that humans pronounce pseudowords by a rule-based process (e.g., Forster & Chambers, 1973). However, Glushko (1979, 1981) has proposed an alternative to the ruled-based theory that does not require the generation of explicit spelling-to-sound rules (see also Brooks, 1977; Baron, 1977). Instead, Glushko has suggested that humans use a process of analogy to derive the pronunciation for a spelling pattern. Words that are similar in spelling to a pseudoword are activated in the language user's mental lexicon. The activated phonological representations in the lexicon are then combined to form an appropriate pronunciation for the novel string.

Glushko (1979) presented a number of experimental results that support his hypothesis. In one study, he found that an "exception pseudoword", that is, a pseudoword that closely resembles words with conflicting pronunciations, will take longer to pronounce than a "regular pseudoword" that resembles a set of words with consistent pronunciations. In addition, he found that words that have inconsistent neighbors (i.e., they resemble words with conflicting pronunciations) are pronounced more slowly than words with consistent

Orthography (Spelling)	Phonetic Representation (Pronunciation)
NILF	/nIlf/
HEEN	/hin/
NICH	/nItS/
POMB	/pam/
LOME	/lom/
HOAP	/hop/
MOOF	/muf/

Figure 1. Example pseudowords that most English speakers can pronounce easily, with their likely pronunciations.

neighbors. Glushko inferred that these exception words and pseudowords are pronounced more slowly than regular strings because the subject must somehow resolve the inconsistency in possible pronunciations.

Examples of regular and exception pseudowords are shown in Figure 2. Thus, T-A-V-E may be pronounced as /tev/ or /tæv/, while T-A-Z-E is typically pronounced as /tez/. Glushko's findings suggest that the pronunciation of both words and pseudowords depends, at least to some degree, on the pronunciation of other words with similar spellings.

Insert Figure 2 about here

The pronunciation-by-analogy theory may provide the same pronunciation ability as a set of spelling-to-sound rules without requiring an explicit theory of rule induction. As a result, for text-to-speech systems, analogy-based pronunciations may eliminate the need for an expert linguist and may be relatively simple to automate. However, pronunciation-by-analogy depends on two critical assumptions. First, the correspondence between spellings and sound patterns in the lexicon is assumed to be sufficiently close to provide reasonable pronunciations. Second, the similarity in spelling patterns between some unkown target letter string and words in the lexicon is assumed to allow the synthesis of a pronuciation from the intersection of spelling-to-sound mappings for several words. To test these assumptions, we developed a computer program called PRONOUNCE that automatically generates a set of rank-ordered pronunciations, in the form of a sequence of phonetic segments, by analogy with the words in a large lexical database. In the sections below, we describe the main features of PRONOUNCE and then summarize its performance in pronouncing novel strings.

II. PRONOUNCE

PRONOUNCE was written in Zetalisp on a Symbolics 3670 Lisp machine. PRONOUNCE contains four basic components. In addition to the lexical database, these include the matcher, which compares the target spelling pattern to the words in the database, the pronunciation lattice, which is a data structure representing possible pronunciations, and the decision function, which rank orders pronunciations extracted from the lattice.

The lexical database used by PRONOUNCE consists of approximately 20,000 words based on Webster's Pocket Dictionary. Each entry includes a mapping from the letters of the word onto the phonetic segments of its pronunciation. This mapping was carried out by a simple Lisp program that only uses knowledge about which letters and phonemes are consonants and which letters and phonemes are vowels. The program parses spellings and pronunciations into separate groups of consonants and vowels, then maps consonant spelling groups to consonant phoneme groups, and vowel spelling groups to vowel phoneme groups.

Insert Figure 3 about here

	Spelling	Pronunciation
Exception:	TAVE	/tæv/ or /tev/
Neighbors:	HAVE	/hæv/
	GAVE	/gev/
Regular:	TAZE	/tez/
Neighbors:	DAZE	/dez/
	GAZE	/gez/

Figure 2. Exception pseudoword TAVE and regular pseudoword TAZE, with their lexical neighbors and pronunciations.

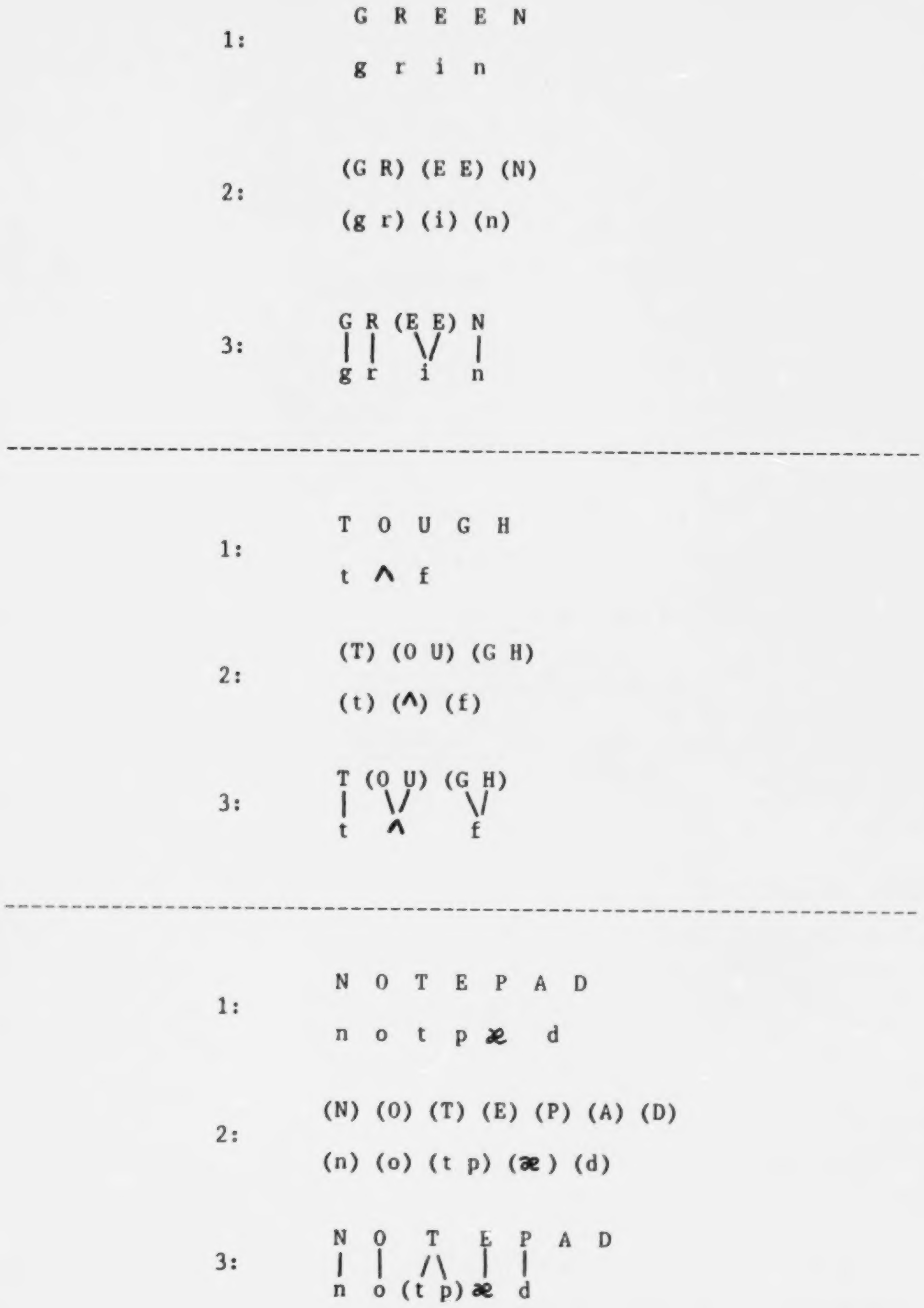

1: G R E E N
g r i n
2: (G R) (E E) (N)
(g r) (i) (n)
3: G R (E E) N
g r i n
1: T O U G H
t ʌ f
2: (T) (O U) (G H)
(t) (ʌ) (f)
3: T (O U) (G H)
t ʌ f
1: N O T E P A D
n o t p æ d
2: (N) (O) (T) (E) (P) (A) (D)
(n) (o) (t p) (æ) (d)
3: N O T E P A D
n o (t p) æ d

Figure 3. The spelling-to-sound mapping process as applied to the words GREEN, TOUGH, and NOTEPAD. In each panel, the spelling and pronunciation of the word is shown at the three stages of the mapping process: 1) before mapping, 2) after parsing into consonant/vowel groups, and 3) after congruent groups have been mapped to each other in left-to-right fashion. The bottom panel shows an incorrectly mapped word.

Figure 3 shows an example of the mappings that would be generated for the words "green", "tough", and "notepad". The mappings produced generally conform to linguistic intuitions, although there are some errors. For example, a "silent e" sometimes occurs between two consonants in a compound word, as shown for "notepad" in the bottom panel of Figure 3. In cases such as this, the "e" is incorrectly mapped onto the next phonetic vowel. This problem could be solved by performing a nonlinear warping of the spelling and phone strings starting with the initial and final segments of both strings and working toward the middle (see Sankoff & Kruskal, 1983).

For each spelling pattern that must be converted to a phoneme pattern, PRONOUNCE looks at every word in the lexicon. The spelling of each lexical entry is compared with the input string by aligning the strings from the beginning, and then sliding the shorter string to the right by one position at a time until the ends of the strings are aligned. For each alignment position, the phonetic segments corresponding to the correctly matched substrings within the word are entered into the pool of information used in generating a pronunciation.

Insert Figure 4 about here

Figure 4 shows an example of the matching process that compares the spelling patterns. In this figure, the input pseudoword "blope" is matched against the lexical entry "sloping". At the first alignment position, the substring "lop" is matched. Subsequent alignment positions do not yield any more matching substrings. Following this spelling pattern matching process, the phonetic mappings are retrieved for the matching substring. The data structure used to represent the output of the matching process as it is applied to the lexical database is a pronunciation lattice. This lattice consists of a set of nodes and arcs. Each node represents a hypothesis about a phonetic segment that may occur at a particular position within the pronunciation. When a letter substring from a lexical entry matches a part of the input string, nodes are created for each phonetic segment produced from the substring. Each of these nodes is connected by arcs to subsequent nodes. In addition to the phoneme nodes, there is a START node at position zero, and an END node at the position that is one greater than the length of the input string. These special nodes mark the entry and exit points for the pronunciation paths in the lattice. Each complete path through the lattice, from START to END, represents a possible pronunciation for the target string.

Insert Figure 5 about here

Figure 5 shows the partial pronunciation lattice that is produced by matching "blope" to "sloping". The correctly matched substring "lop" yields the nodes labeled l-2, o-3, and p-4. In this example, the node labeled "l-2" asserts that the phonetic segment corresponding to the second letter in the pseudoword is /l/. Node l-2 is connected to o-3, representing the fact that a string "lo" was matched, resulting in the phonetic sequence /l o/. Nodes o-3 and p-4 are similarly connected. In addition, node l-2 is connected directly to node p-4, representing the match for the complete substring "lop".

```
                INPUT:  BLOPE

           LEXICAL ENTRY:  SLOPING



-------------------------------------------------------------------

                1st position:

                B L O P E
                S L O P I N G

        shared:   L O P

-------------------------------------------------------------------

                2nd position:

                  B L O P E
                S L O P I N G

        shared:  none

-------------------------------------------------------------------

                3rd position:

                    B L O P E
                S L O P I N G

        shared:  none
```

Figure 4. An example of the matching process that finds spelling pattern analogies. In this example, the input pseudoword "blope" is matched against the lexical entry "sloping" at three orthographic alignment positions.

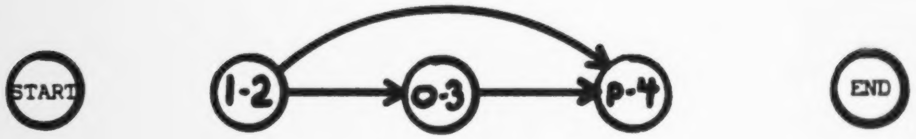

START
l-2
o-3
p-4
END

Figure 5. A segment of the pronunciation lattice for the "lop" substring.

Attached to the arc connecting 1-2 and p-4 is a list of phonemes to be added to the pronunciation should that arc be taken. In this case the list consists of the phoneme /o/.

Also, a frequency count is stored with each arc, reflecting the number of matched substrings that produced that arc. Finally, if a matched substring starts at the beginning of a word or ends at the end of a word, the START and END nodes are connected to the appropriate nodes of the pronunciation lattice segment.

Once the entire lexicon has been processed and the complete lattice has been constructed, the lattice is traversed to find all the paths from START to END. These paths represent all the pronunciations for the input string. As the lattice is traversed, PRONOUNCE keeps track of the number of arcs in each path, as well as the sum of the arc frequencies for each path. The paths are then rank-ordered, first by length, and second, by the sum of the arc frequencies. By using path length as a heuristic for selecting pronunciations, decisions are based on the amount of spelling overlap between an input string and the lexicon. The use of arc frequencies as a secondary heuristic causes the system to select the most common spelling-to-sound translations that occur in the database.

III. RESULTS

We tested PRONOUNCE on 70 pseudowords presented by Glushko (1979) to human subjects for pronunciation. All the strings were four or five characters in length and they were all derived from monosyllabic words by changing one letter. Since these pseudowords have no objectively correct pronunciation, we also elicited pronunciations from human subjects, to give us a basis for assessing the performance of PRONOUNCE. The seven subjects all had formal training and experience in phonetic transcriptions of English words. The subjects were given a printed list of the pseudowords and were instructed to provide a phonetic transcription of the first pronunciation that came to mind. In order to compare the performance of PRONOUNCE with a system using a set of relatively complex spelling-to-sound rules, we also presented the pseudowords in ASCII format to Digital Equipment Corporation's DECtalk V2.0 text-to-speech system, which provided phonetic transcriptions.

Insert Figure 6 about here

Figure 6 shows some sample pronunciations produced by the human subjects, PRONOUNCE, and DECtalk. In general, both PRONOUNCE and DECtalk consistently agree with the pronunciation given by the human subjects.

A pronunciation was scored correct if it exactly matched one of the pronunciations produced by one of the human subjects. PRONOUNCE performed with an error rate of 9%, while DECtalk had an error rate of 3%.

IV. SUMMARY AND CONCLUSIONS

The success of the PRONOUNCE system provides some additional support for the hypothesis that humans pronounce novel words and pseudowords by analogy to known words in their mental lexicons. It is clear that pronunciation-by-analogy is computationally sufficient to generate reasonable pronunciations of

Pseudoword	Human	PRONOUNCE	DECtalk
COTH	/kaθ/	/kaθ/	/kaθ/
HEEN	/hin/	/hin/	/hin/
DROOD	/drud/	/drud/	/drud/
SHEAD	/šid/	/šid/	/šid/
FEAD	/fid/	/fɛd/	/fid/
STEAT	/stit/	/stIt/	/stid/
POMB	/pam/	/pam/	/pam/
COSE	/kos/ /koz/	/kos/	/koz/

Figure 6. Comparison of sample human, PRONOUNCE, and DECtalk transcriptions.

novel words. Thus, pronunciation-by-analogy can account for the performance of humans, even though the performance of PRONOUNCE must still be improved to completely simulate human performance. Moreover, PRONOUNCE demonstrates that pronunciation-by-analogy may, with sufficient development effort, replace the use of spelling-to-sound rules in the next generation of text-to-speech systems. This will reduce the need for human intervention in modifying a text-to-speech system for a new dialect or language.

In addition to these considerations, pronunciation-by-analogy may improve spelling-to-sound translation for surnames (cf. Church, 1985; Spiegel, 1985). Surnames are often "borrowed" into English from other languages, and the application of English spelling-to-sound rules to these names often produces inappropriate translations. However, with some representative entries for foreign names in the lexical database, pronunciation-by-analogy should be able to produce the correct spelling-to-sound translation for other names from the sampled language.

Finally, it is clear that additional research efforts on pronunciation-by-analogy are needed. First, the present system of mapping letter strings to phoneme strings could be modified to incorporate non-linear string warping and more linguistic knowledge. At the present time, only the distinction between consonants and vowels is used to bind phoneme translations onto letter substrings. This mapping process could be modified to incorporate more sophisticated representations of syllable structure and stress pattern, and therefore generate more linguistically complex mappings. Second, the string-matching algorithm represents the core of the analogy process in the present version of PRONOUNCE. Improvements in this analogy function should result in selecting lexical entries that are more appropriate for building the pronunciations of an input string. In the present version, a match of a single letter between the input string and a lexical entry causes phonetic information to be added from the lexical entry to the pronunciation lattice. In the next version of PRONOUNCE, we are planning to incorporate a more abstract concept of analogy based on the overall patterns of spelling information in the input and lexical entry and a more sophisticated way of computing similarity or distance scores for two strings.

In summary, PRONOUNCE demonstrates that pronunciation-by-analogy may provide a viable alternative to traditional rule-based pronunciation systems. However, improvements in the pronunciation generated by analogy will depend on incorporating more explicit linguistic knowledge and more abstract concepts of analogy into the system. Finally, there is a need for more research on the structural properties of spelling and sound patterns in order to develop more sophisticated theories of pronunciation by humans and eventually to incorporate these more powerful and productive systems into the next generation of text-to-speech converters.

References

Baron, J. (1977). Mechanisms for pronouncing printed words: Use and acquisition. In D. LaBerge and S. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, NJ: Erlbaum.

Brooks, L. (1977). Non-analytic correspondences and pattern in word pronunciation. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.

Church, K. (1985). Stress assignment in letter to sound rules for speech synthesis. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, July, Chicago.

Forster, K. I., and Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627-635.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674-691.

Glushko, R. J. (1981). Principles for pronouncing print: The psychology of phonography. In A. M. Lesgold and C. A. Perfetti (Eds.), *Interactive processes in reading*. Hillsdale, NJ: Erlbaum.

Sankoff, D., and Kruskal, J. B. (1983). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Reading, MA: Addison-Wesley.

Spiegel, M. F. (1985). Pronouncing surnames automatically. *Proceedings of the Voice I/O Systems Applications Conference '85*, September, San Francisco.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

The Role of the Lexicon in Speech Perception*

David B. Pisoni, Paul A. Luce, and Howard C. Nusbaum

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

*This is the final draft of a paper prepared for the NATO Advanced Research Workshop "The Psychophysics of Speech Perception," held at Utrecht University, The Netherlands, June 30--July 4, 1986. This research was supported, in part, by NIH Research Grant NS-12179 to Indiana University in Bloomington. Preparation of this manuscript was supported by an award from the James McKeen Cattell fund to David B. Pisoni.

The Role of the Lexicon in Speech Perception

For the last five years our research group at Indiana has been interested in the role of the lexicon in speech perception. To carry out this work, we have adopted a computational approach that makes use of several large databases containing phonetic transcriptions and appropriate software to index the structural properties of words and relations among words. Our initial work has revealed a number of interesting and potentially important findings about the distribution of words in the mental lexicon and the role of structural constraints in human word recognition (see Pisoni, Nusbaum, Luce & Slowiaczek, 1985). We believe this computational approach to word recognition may provide insights into several very long-standing problems in the field of human word recognition and may have some important implications for large vocabulary speech recognition in terms of developing computational techniques for rapid and efficient search space reduction. Although there has been a great deal of research carried out over the last few years on the development of metrics for quantifying spectral distance among speech signals, there has been little, if any, research directed at formalizing an approach to phonetic distance among words in terms of their acoustic-phonetic or phonological similarity. Moreover, there has been practically no research done to determine if these proposed distance metrics based on the structural properties and similarity spaces of words can be used to predict the behavior of human listeners. Finally, there has been little, if any, knowledge of the precise role that structural constraints play in speech perception and word recognition.

The overall goal of our research on the lexicon is to learn more about the structural properties of words in the mental lexicon and how this information may be used by human observers in recognizing words. We are interested in learning how structural information can be used to reduce the search space and how it constrains the selection of lexical candidates in the word hypothesization process. We have been interested in determining which aspects of the structural properties of words are important in this task and how this information may be used by human listeners in recognizing spoken words from very large vocabularies. As a side interest, we have also been considering how information derived from studies of human speech perception and word recognition could be used in developing new and more efficient algorithms for speech recognition. However, our major interest is in human auditory word recognition and speech perception.

As a first step to approaching these computational problems, we acquired several large computer readable databases. One of these, based on Kenyon and Knott's A Pronouncing Dictionary of American English and Webster's Seventh Collegiate Dictionary, contains approximately 300,000 entries. Another smaller database of 20,000 words is based on Webster's Pocket Dictionary. Each entry contains the standard orthography of a word, several phonetic transcriptions in ARPABET, stress patterns and special codes indicating the syntactic functions of the word. We have developed a number of algorithms for determining, in various ways, the similarity neighborhoods, or "lexical density," for any given entry in the dictionary based on analyses of either the orthographic or phonetic transcriptions of words in the database. This information has provided some very interesting information about the structural properties of words in the lexicon and how this information might be used by human listeners in word recognition. In the sections below, we summarize briefly some of our recent findings using these computational techniques and discuss their implications for spoken word recognition.

Word Frequency Effects. Word frequency effects in the psychological literature obtained in perceptual and memory research using human subjects have typically been accounted for in terms of frequency of usage, the time between the current and last encounter with the word in question, and similar ideas. In each of these explanations of word frequency effects, however, it has been assumed that high and low frequency words are "perceptually equivalent." That is, it has often been assumed that common and rare words are structurally equivalent in terms of phonological and orthographic composition. A number of years ago, Landauer and Streeter (1973) suggested that the assumption of perceptual equivalence of high and low frequency words may not necessarily be warranted. In their study, Landauer and Streeter demonstrated that common and rare words differ on two structural dimensions. From an analysis of a small number of printed words, they found that the "similarity neighborhoods" of common and rare word differ in both size and composition: High frequency words have more words in common (in terms of one letter substitutions) than low frequency words, and high frequency words tend to have high frequency neighbors, whereas low frequency words tend to have low frequency neighbors. Thus, for printed words, the similarity neighborhoods for high and low frequency words show marked differences. Landauer and Streeter also demonstrated for a small number of spoken words that certain phonemes are more prevalent in high frequency words than in low frequency words and vice versa. Both of these analyses and conclusions were, however, based on a very small number of words and no attempt was ever made to generalize these findings to a larger database that is more representative of the words in the language.

In our laboratory, Luce has recently undertaken a project that is aimed at extending and elaborating the original Landauer and Streeter study (Luce, 1986). In this research, both the similarity neighborhoods and phonemic constituencies of high and low frequency words have been examined in order to determine the extent to which spoken common and rare words differ in the nature and number of "neighbors" as well as their phonological configuration. To address these issues, an on-line version of *Webster's Pocket Dictionary* (WPD) was employed to compute statistics about the structural organization of words. Specifically, the phonetic representations of approximately 20,000 words were used to compute similarity neighborhoods and to examine phoneme distributions for words of different frequencies of occurrence in the language.

To examine the similarity neighborhoods of common and rare words, a subset of high and low frequency target words were selected from the WPD for evaluation. High frequency words were defined as those equal to or exceeding 1000 words per million in the Kucera and Francis word count. Low frequency words were defined as those between 10 and 30 per million inclusively. For each target word meeting these a priori criteria, similarity neighborhoods were computed based on a metric allowing only one-phoneme substitutions at each position within the target word. There were 92 high frequency words and 2063 low frequency words. The mean number of words within the similarity neighborhoods for the high and low frequency words was also computed, as well as the mean frequencies of the neighbors. In addition, a decision rule was used to was compute a measure of the "distinctiveness" of a given target word relative to its neighborhood with the following formula:

$$\frac{T}{T + \sum N_i}$$

where T equals the frequency of the target word and N equals the frequency of the i-th neighbor of that target word. Larger values for the decision rule indicate a target word that "stands out" in its neighborhood; smaller values indicate a target word that is relatively less distinctive in its neighborhood.

The results of these analyses revealed that although the mean number of neighbors for high and low frequency target words were approximately equal, the mean frequencies of the *similarity neighborhoods* for high frequency target words were higher than the mean frequencies of the similarity neighborhoods of the low frequency target words. The finding that high frequency words tend to have neighbors of higher frequency than low frequency words suggests, somewhat paradoxically, that high frequency words are *more likely* rather than less likely to be confused with other words than low frequency words.

At first glance, this finding would appear to contradict the results of many behavioral studies reported in the psychological literature demonstrating that high frequency words are recognized more easily than low frequency words. However, an examination of the predictions derived from the decision rule applied to high and low frequency target words showed that high frequency words should be perceptually distinctive relative to the words in their neighborhoods whereas low frequency targets should not. Indeed, substantially larger values of this index were obtained for high frequency words than for low frequency words of the same length.

Several other interesting findings were also revealed in these analyses. First, for target words of both high and low frequencies, the decision rule predicted increasingly better performance for words of greater length. In addition, the analyses showed that for words consisting of more than three phonemes, the percentage of unique words increased very substantially as word length increased. This last finding demonstrates that simply increasing the length of a word increases the probability that the phonotactic configuration of that word will be unique and eventually diverge from all other words in the lexicon. Such a result suggests the potentially powerful contribution of word length in phonemes in combination with various structural factors to the isolation of a given target word in the lexicon.

Phoneme Distributions in High and Low Frequency Words. The finding that high frequency spoken words tend to be more similar to other high frequency words than to low frequency words also suggests that certain phonemes or phonotactic configurations may be more common in high frequency words than in low frequency words. As a first attempt to evaluate this claim, Luce has examined the distribution of phonemes in words having frequencies of 100 or greater and words having a frequency of one. For each of the 45 phonemes used in the transcriptions contained in the WPD, percentages of the total number of possible phonemes for four and five phoneme words were computed for the high and low frequency subsets.

Of the numerous trends uncovered through these analyses, two were of special interest. First, the percentages of bilabials, interdentals, palatals, and labiodentals tended to remain relatively constant or decrease slightly from the low to high frequency words. However, the pattern of results for the alveolars and velars was quite different. For the alveolars, increases from low to high frequency words of 9.07% for the four phoneme words and 3.63% for the five phoneme words were observed. For the velars, the percentage of phonemes dropped from the low to high frequency words by 2.33% and 1.14% for the four and five phoneme words, respectively. Second, there was an increase of 4.84% for the nasals from low to high frequency words accompanied by a corresponding drop of 4.38% in the overall percentage of stops for the five phoneme words.

The finding that high frequency words apparently tend to favor consonants having an alveolar place of articulation and disfavor those having a velar place of articulation suggests that frequently used words in the language may have succumbed to pressures over the history of the language to exploit consonants that are, in some sense, "easier" to articulate for human talkers. This result, taken together with the finding for five phoneme words regarding the differential use of nasals and stops in common and rare words, suggests that, in terms of phonological composition, common words differ structurally from rare words in terms of their choice or selection of constituent phonemes. Further analyses of the phonotactic configuration of high and low freqeuncy words should reveal even more striking structural differences between high and low frequency words in light of the results obtained from the crude measure of structural differences based on the overall distributions of phonemes in these words.

Similarity Neighborhoods and Word Identification. In addition to the work summarized above demonstrating differences in structural characteristics of common and rare words, we have also explored the use of similarity neighborhoods as a measure of lexical density to derive predictions regarding word intelligibility in noise. A subset of 300 words published by Hood and Poole (1980) which were ranked according to their intelligibility in white noise have been examined to study the role of similarity neighborhoods in word recognition. As Hood and Poole pointed out, frequency of usage was not consistently correlated with word intelligibility scores in their data. In our analyses, we reasoned that some metric based on the similarity neighborhoods of these words might be better at capturing the observed differences in intelligibility than a simple account based only on frequency of occurrence in the language.

To test this possibility, we examined 50 of the words provided by Hood and Poole, 25 of which constituted the easiest words and 25 of which constituted the most difficult words in their data. In keeping with Hood and Poole's observation regarding the effects of word frequency, we found that the 25 easiest and 25 most difficult words were not, in fact, significantly different from each other in frequency. However, we found that the relationship of easy words to their neighbors differed very substantially from the relationship of the difficult words to their neighbors. More specifically, on the average, 56.41% of the words in the neighborhoods of the difficult words were equal to or higher in frequency than the difficult words themselves, whereas only 23.62% of the neighbors of the easy words were of equal or higher frequency. Thus, it appears that the observed differences in intelligibility of these words may have been due, at least in part, to the frequency composition of the neighborhoods of the easy and difficult words, and were obviously not primarily due to the absolute frequencies of the words themselves (Anderson, 1962; Havens and Foote, 1963). Thus, it appears that the difficult words found in Hood and Poole's study were more difficult to perceive because they had relatively more "competition" from their neighbors than the easy words.

Phonotactic _Patterns_ _of_ _Words_ _in_ _the_ _Lexicon_. Although human listeners may perceive spoken words as a temporally distributed sequence of segments, the recognition system need not compare these segments to lexical representations in memory in a strict left-to-right order as assumed by several current theories of word recognition. Indeed, it is unclear how serial pattern matching strategies can recognize a word if the initial segment of the input is somehow obscured, degraded or ambiguous. Because this initial segment may be thought of as an index into the lexicon, as in a content addressable memory system (Kohonan, 1980), recognition could not proceed without a well-defined access point to begin the process. This is obviously one of the reasons why the problem of segmentation has been so important in automatic speech recognition. An alternative approach is to view auditory word recognition as process involving "constraint satisfaction" rather than simply pattern recognition of elementary features or attributes. According to this view of word recognition, the propagation of a number of weak constraints is used to specify the target word. When word recognition is viewed as a process of constraint satisfaction, a number of quite different sources of information can be simultaneously applied to the lexicon in parallel to refine the set of hypothesized word candidates. Even if one constraint is uninformative, the intersection of the other constraints across different domains may still be able to specify the correct word and locate it even in very large search spaces. Given this view, it is important to determine precisely which constraints, if any, may be used by human listeners in auditory word recognition.

The approach we have taken to study the role of structural constraints in auditory word recognition was motivated, in part, by several recent studies that investigated the relative heuristic power of several different classification schemes for large vocabulary word recognition by computers. Zue and his colleagues (Huttenlocher & Zue, 1984; Shipman & Zue, 1982) have shown that a _partial_ phonetic specification of _every_ phoneme in a word results in an average candidate set size of about two words for a vocabulary of 20,000 words. The partial phonetic specification in Zue's system consisted of six broad manner classes. Thus, with this approach, a recognition system need not accurately identify the exact phonemes in a spoken word. Instead, only the most robust manner information needs to be coded. Using a slightly different approach, Crystal et al. (1977) demonstrated that increasing the "phonetic refinement" of every phoneme in a word from four broad phonetic categories to ten more refined categories produced large improvement in the number of unique words identified in a large corpus of text.

It is important to point out here that these computational studies examined the consequences of partially classifying _every_ segment in a word. Thus, two constraints were actually employed: the partial classification of each segment in the word _and_ the broad phonotactic shape of each word resulting from the combination of word length with coarse phonetic information about each segment. The analyses that we have carried out used a large lexical database consisting of a subset of 126,000 words from the Kenyon and Knott pronouncing dictionary. In carrying out these analyses, we assumed that human listeners are able to recognize much more phonetic information than would be encoded by just six coarse manner categories. Thus, in carrying out our analyses, we assumed that human listeners might be able to identify some segments completely, other segments only partially and finally some segments not at all.

The results of our analyses have been quite revealing about the structural constraints that may contribute to search space reduction in auditory word recognition. For the most gross level of segmental analysis, that is, knowing only the length of a word in terms of the number of constituent phonemes, the search space was reduced from 126,000 words to only 6,342. Thus, word length per se appears to be an extremely powerful constraint for reducing the candidate set in the lexicon by at least two orders of magnitude, even without knowing any detailed segmental phonetic information about the internal structure of the word. Our analyses also revealed that this length constraint is strongest for the longest words. Thus, as words become longer and longer, less detailed segmental information is needed to isolate the hypothesized candidates in recognition.

By simply classifying each segment in a word as either a consonant or vowel, without providing any more detailed phonetic information, the reduction in the search space beyond the length constraint becomes even more substantial. The number of candidates is reduced by another order of magnitude to 109 words averaged across different word lengths. It is interesting to note here that much of this reduction in the size of the candidate set appears to be due to the specific phonotactic constraints imposed by the ordering of consonants and vowels in words. If the segments in a word are classified with just two categories, consonants and vowels, but the ordering of these categories is removed, then the average candidate set size increases to 1196 words instead of the 109 words observed earlier. This finding suggests that the phonotactic information in the pattern structure of a spoken word accounts for another order of magnitude reduction in the candidate search space compared with just having information available about the number of consonants and vowels within a word or its length.

Increasing the amount of phonetic detail for each segment from two categories (i.e., consonants and vowels) to six coarse manner classes used by Zue and his colleagues, reduces the search space by another two orders of magnitude from the CV classification scheme. The average candidate set size is reduced to about 5.5 words from the original 126,000 words. Our analyses of words from the 126,000 lexicon agree closely with the findings reported by Shipman and Zue (1982) using the 20,000 word Webster's Pocket Dictionary. Increasing the size of the lexicon by an order of magnitude from 20,000 words to 126,000 words only results in a tripling of the number of lexical candidates from 2 to about 6 words. Thus, by any metric, partial information about every segment is an extremely powerful constraint on the candidate set of words.

Of course, research over the last 80 years on human word recognition has shown that listeners are able to resolve much more phonetic detail in the speech waveform than just six broad manner categories. One issue that we considered concerns the constraint that is provided by complete phonetic information about some of the segments in a word compared to only partial information about every segment in a word. Classifying every segment in a word provides two types of information: (1) partial phonetic information about every segment, and (2) the phonotactic "shape" or envelope of the entire word. By comparison, complete phonetic classification of only some of the segments in a word provides: (1) detailed phonetic information about a few segments, and (2) partial information about the phonotactic shape of a word. Based on the previous demonstrations of the power of the phonotactic shape of a word with only two categories corresponding to consonant and vowel, it seems reasonable to predict that a partial classification of every segment in a word should be more effective than complete classification of some of the segments.

To test this prediction, we carried out the following analyses. First, the phonetic information in the first half of every word was classified very narrowly leaving the remaining segments unclassified with cover symbols.

Second, the phonetic information in the last half of each word was classified very narrowly therefore leaving the first half of each word unclassified. Two other analyses were carried out in which only the consonants or only the vowels were classified completely. When the consonants were classified, the vowels were left unspecified and vice versa. To our surprise, given the earlier prediction, the results showed that complete information about only some of the segments in a word actually provides a more powerful constraint on search space reduction than having partial information about every segment. Classifying the beginning of words completely reduces the search space from 126,000 words to 1.7 words and classifying the last half of words reduces the candidate set to only 1.9 words. By comparison, classifying only the consonants exactly and leaving the vowels unspecified yielded a candidate set size of 1.4 words whereas classifying the vowels and leaving the consonants unspecified yielded a set size of 3.2 words.

In short, complete phonetic information about some of the segments in a word constrains the search space much more than partial classification about every segment. Our computational analyses suggest that detailed phonetic information about some of the segments in a word provides enough constraint, in general, so that other segments can probably be obscured or remain ambiguous without significantly impairing recognition. Moreover, to the extent that some phonetic information is available about other segments, the candidate set will be reduced even further, probably to the extent of uniquely specifying the correct word or one that is highly similar to it. Of course, these analyses are based on computations carried out with a large database and it remains to be seen if these same findings generalize to how humans listeners recognize words from large search spaces. This work is currently underway.

In summary, the results obtained thus far using a number of computational techniques with databases of phonetic transcriptions suggest that the processes involved in word recognition may be highly contingent on the structural factors related to the organization of words in the lexicon and the relation of words to other phonetically similar words in surrounding neighborhoods in the lexicon. The outcome of this work should prove quite useful in discovering not only the underlying structure of the sound patterns of words in the mental lexicon, but also in detailing the implications these structural constraints may have for the real-time processing of spoken language by human listeners as well as machines. In the case of machine recognition of speech, our findings may provide a principled way to develop new distance metrics based on acoustic-phonetic similarity of words in large vocabularies that could substantially reduce the search space for lexical hypothesization.

The present research is concerned with a central problem in large vocabulary word recognition, namely, the organization of the sound patterns of words and the structural constraints that define these patterns in spoken language. We view the lexicon as a complex multidimensional space in which phonetically similar sound patterns are grouped more closely together than phonetically dissimilar patterns. A major task in our future research is to define the dimensions of this similarity space and specify the structural constraints that are used to characterize the similarity neighborhoods of words in this space. With these findings, we should be able to predict the behavior and error patterns of human observers in large vocabulary word recognition tasks and contribute new knowledge to the development of more efficient algorithms for large vocabulary search space reduction in automatic speech recognition. We believe that the current conception of the mental lexicon as simply a large dictionary with words indexed by frequency of occurrence in the language is inadequate to account for many of the behavioral findings reported in the literature. Moreover, our preliminary findings suggest a potentially more powerful approach in terms of conceptualizing the lexicon in terms of similarity spaces for sound patterns of words in terms of neighborhoods. Thus, in our view, words should not be conceived of as unrelated lexical entries as in a dictionary but rather as auditory patterns that have a complex internal structure reflecting the morphology, phonology and phonotactics of English. The process of recognizing a word should therefore not be viewed as a form of pattern recognition via template matching or feature analysis but rather as an ideal example of recognition via constraint satisfaction.

References

Anderson, D. C. (1962) The number and nature of alternatives as an index of intelligibility. Unpublished doctoral dissertation, Ohio State University.

Greenberg, J. H. and Jenkins, J. J. (1964) Studies in the psychological correlates of the sound system of American English. Word, 20, 157-177.

Havens, L. L. and Foote, W. E. (1963) The effect of competition on visual duration threshold and its independence of stimulus frequency. Journal of Experimental Psychology, 65, 6-11.

Hood, J. D. and Poole, J. P. (1980) Influence of the speaker and other factors affecting speech intelligibility. Audiology, 19, 434-455.

Kohonen, T. (1980) Content-Addressable Memories. NY: Springer-Verlag.

Landauer, T. K. and Streeter, L. A. (1973) Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior, 12, 119-131.

Luce, P. A. (1986) Structural distinctions between high and low frequency words in auditory word recognition. Unpublished doctoral dissertation, Indiana University.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A. and Sloviaczek, L. M. (1985) Speech perception, word recognition and the structure of the lexicon. Speech Communication, 4, 75-95.

The Role of Structural Constraints in Auditory Word Recognition*

Howard C. Nusbaum and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

*This research was supported in part by NIH Research Grant NS-12179 to Indiana University in Bloomington. This paper was presented at the Montreal Symposium on Speech Recognition, Montreal, Canada, July, 1986. A version of this paper appears in the proceedings of that symposium.

Abstract

Auditory word recognition in humans has traditionally been viewed as a pattern matching process in which acoustic-pattern attributes of a spoken word are compared to stored representations of words in the mental lexicon. One consequence of this view is that each word is treated as an isolated pattern that is independent of the other words in the lexicon. An alternative approach to understanding human auditory word recognition is to consider recognition as a search process that takes place within a lexical space defined by the pattern structures of words in the listener's vocabulary. By this view, word recognition becomes a process of constraint satisfaction for which the constraints are the acoustic-pattern attributes of a spoken word and the structural relationships among words in the mental lexicon. In this paper, we consider some of the implications of this view of human auditory word recognition.

The Role of Structural Constraints in Auditory Word Recognition

In the past, much of the research on human speech perception has focused on the recognition of acoustic-phonetic properties of isolated CV and CVC syllables. The tacit assumption of this research has been that our understanding of auditory word recognition is contingent upon solving the problems inherent in phoneme perception. By this assumption, auditory word recognition is equivalent to visual word recognition carried out one letter at a time. Indeed, most current theories of auditory word recognition directly reflect this sequential pattern matching approach to word recognition. However, a different perspective is that word perception may be approached as a problem of "weak" constraint satisfaction, in which the structural properties of words in the lexicon interact to specify the identity of an utterance. We will present the results of several analyses of the phonotactic constraints of word patterns that suggest the type of constraints that may be used by human listeners to mediate spoken word recognition.

Recognition in the Context of the Lexicon

Context exerts an undeniably strong influence on perceptual processes. However, it is interesting to note that "context" is defined in almost all speech research by whatever stimulus information is presented immediately prior to or subsequent to a target stimulus. Thus, a phoneme is perceived in the context of a syllable, a syllable is perceived in the context of a word, and a word is perceived in the context of a sentence. In all cases, there are objectively definable physical dimensions to the context that is typically investigated. But there is another context that affects word perception as well: the implicit context of the mental lexicon. Beyond the listener's explicit knowledge about words, the structure and organization of the sound patterns of lexical entries may serve as an implicit context within which recognition occurs.

Marslen-Wilson and Welsh (1978) called attention to the potential importance of the structural properties of words with the cohort theory of word recognition. According to this theory, the initial sounds in a stimulus word activate all the words in the lexicon beginning with those sounds. Inappropriate candidates in the cohort are then deactivated when a mismatch occurs in comparing the left-to-right order of subsequent segments in the stimulus with the structures of activated candidates. The word that is ultimately recognized is the candidate that remains after all the other incompatible candidates have been deactivated.

According to cohort theory, the activated cohort of word candidates in the lexicon forms the mental context for spoken word recognition. However, unlike the sentential context that may precede a spoken word, this context has no physical dimensions that can be directly measured or analyzed. In the past, this has posed a problem for investigating the role of the lexicon in word recognition. However, several computer-readable databases of orthographic and phonetic representations of words have recently become available for analyzing the structural properties of words in the lexicon. The database used for all the analyses we will describe contains orthographic, phonetic, and syntactic information for 243,000 words (see Crystal, Hoffman, & House, 1977). Proper names and possessives were excluded from the analyses, leaving about 126,000 words that were examined in the database.

Phonotactic Patterns in the Lexicon

Although the listener may be presented with spoken words as a temporally distributed sequence of segments, a recognition process need not compare these segments to lexical representations in a strict left-to-right order as claimed by some theories. Indeed, it is unclear how serial pattern matching strategies can recognize a word if the initial segment of the input is obscured, degraded or ambiguous. Since this initial segment is treated as the index into the lexicon, recognition could not proceed without a well-defined access point. An alternative approach is to view auditory word recognition as a constraint satisfaction process, in which the propagation of a number of weak constraints is used to specify the recognized word. When viewed as a constraint satisfaction process, a number of constraints may simultaneously be applied to the lexicon to refine the set of word candidates. Even if one constraint is inappropriate or uninformative, the intersection of the other constraints may still specify the correct word. Given this view, it is important to determine precisely which constraints are actually used during word perception.

The approach that we have taken to investigate structural constraints on human auditory word recognition was motivated by several recent studies that investigated the relative heuristic power of various classification schemes for large vocabulary word recognition by computers. Zue and his colleagues (Huttenlocher & Zue, 1984; Shipman & Zue, 1982) have shown that a *partial* phonetic specification of *every* phoneme in a word results in an average candidate set size of about 2 words for a vocabulary of 20,000 words. The partial phonetic specification consisted of six broad phonetic manner classes. Thus, with this approach, a recognition system need not accurately identify the phonemes in spoken words. Instead, only the most robust manner information must be coded. Using a slightly different approach, Crystal et al. (1977) demonstrated that increasing the phonetic refinement of every phoneme in a word from four broad phonetic categories to ten more refined categories produces large improvements in the number of unique words identified in a large corpus of text.

It is important to note that these computational studies examined the consequences of partially classifying *every* segment in a word. Thus, they actually employed two constraints: the partial classification of each segment *and* the broad phonotactic shape of each word resulting from the combination of word length with patterned phonetic information.

The analyses that we have carried out used a large lexical database of 126,000 words to study different constraints that might be appropriate for describing human auditory word recognition. This work extends the previous research of Zue and his colleagues to a much larger set of words. In addition, since human listeners are capable of recognizing much more phonetic information than just six manner categories, we have carried out analyses based on the assumption that human listeners will be able to identify some segments completely, while other segments will be unanalyzed.

The results of these analyses are quite revealing about the recognition constraints provided by the structural properties of spoken words. For the coarsest level of segmental analysis, that is, knowing only the length of a word in number of phonemes, the search space is reduced from 126,000 words to 6,342 words. Clearly, word length is a very powerful constraint for reducing the candidate set in the lexicon by about two orders of magnitude, even without any detailed segmental phonetic information. Furthermore, the length constraint is strongest for relatively long words. If the length of a word is 21 segments, there are only two candidates out of 126,000 words. Thus, as word length becomes extreme, less detailed segmental information is needed to identify a word.

By simply classifying each segment as either a consonant or vowel (i.e., two categories), without providing any more detailed phonetic description, the reduction in the search space beyond the the length constraint phonotactic constraint is enormous. The number of candidates is reduced by an order of magnitude to 109 words averaged across different word lengths. Furthermore, it is interesting to note that much of this reduction in the candidate set is due to the specific phonotactic constraints provided by the ordering of consonants and vowels. If the segments in a word are classified with just two categories, as consonants or vowels, but the order information is removed, there are 1196 words in the average candidate set. This means that the phonotactic order information in the pattern structure of a spoken word accounts for an order of magnitude reduction in the candidate set size compared to just knowing the number of consonants and vowels, but not their arrangement.

Increasing the amount of phonetic detail for each segment to the six manner classes used by Zue and his colleagues reduces the search space by another two orders of magnitude from the CV classification scheme that maintains phonotactic order information. Using six categories for classifying every segment in each word reduces the average candidate set size to about 5.5 words from 126,000 words in the lexicon. This result agrees very well with the results reported by Shipman and Zue (1982) for a 20,000 word lexicon, indicating that this broad classification scheme is very powerful in reducing the number of word candidates in the search space. Increasing the lexicon by an order of magnitude from 20,000 words to 126,000 words only results in a tripling of the number of candidates from 2 to about 6 words. By any metric, partial information about every segment is an extremely effective constraint on the candidate set.

However, human listeners are capable of resolving much more phonetic detail than just six broad categories. One issue that can be raised then, concerns the constraint provided by complete phonetic information about some of the segments in a word compared to partial information about every segment in a word. Classifying every segment in a word provides two types of information: (1) partial phonetic information about every segment, and (2) the phonotactic "shape" of the entire word. By comparison, complete classification of some of the segments provides: (1) detailed phonetic information about a few segments, and (2) partial information about the phonotactic shape of a word. Based on the previous demonstration of the power of phonotactic shape with just two categories (i.e., consonant or vowel), it seems reasonable to predict that partial classification of every segment in a word should be more effective than complete classification of some of the segments in a word.

To test this prediction the following analyses were carried out: (1) the phonetic information in the first half of every word was classified completely leaving the remaining segments unclassified, (2) the phonetic information in the last half of each word was classified completely leaving the first half unclassified, (3) only the consonants were phonetically classified leaving the vowels unlabeled, and (4) the vowels were phonetically classified leaving the consonants unlabeled. The results demonstrate that complete information about some of the segments in a word provides a more powerful constraint on the candidate set than partial classification of every segment. Classifying the beginning of words completely reduces the search space from 126,000 words to 1.7 words and classifying the last half of words reduces the candidate set to 1.9 words. By comparison, classifying only the consonants exactly and leaving the vowels unclassified yields a set size of 1.4 words, while classifying the vowels only yields a set size of 3.2 words. In each analyses, complete phonetic information about some of the segments in a word constrains the search space much more than partial classification of every segment. These results demonstrate that detailed phonetic information about some of the segments in a word provides enough constraint, in general, that other segments can be completely obscured or ambiguous without significantly impairing recognition. Moreover, to the extent that some phonetic information is available about other segments, the candidate set will be reduced further, probably to the extent of uniquely specifying the correct word.

Conclusions

The view of word recognition that emerges from these analyses differs substantially from serial pattern matching approaches. As more of a stimulus word is heard, the listener progressively narrows the candidate set based on the development of a phonotactic specification for the input. Over time, acoustic information in the stimulus is successively refined into more detailed phonetic representations. In some cases, only a broad phonetic description of segments may be computable and the phonotactic structure is used to further narrow the candidate set. This approach, called Phonetic Refinement Theory, is currently being implemented as a model of the recognition process. Although further research is needed, it is clear that computational analyses of the sound patterns of words can provide new information about the processes that mediate speech perception.

References

Crystal, T. H. , Hoffman, M. K., & House, A. S. (1977) Statistics of phonetic category representation of speech for application to word recognition. Princeton, NJ: Institute for Defense Analysis.

Huttenlocher, D. P. , & Zue, V. W. (1984) A model of lexical access based on partial phonetic information. *Proceedings of ICASSP-84*, New York: IEEE Press, Volume 2.

Marslen-Wilson, W. D., & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

Shipman, D. W., & Zue, V. W. (1982) Properties of large lexicons: Implications for advanced isolated word recognition systems. *Proceedings of ICASSP-82*, New York: IEEE Press.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

A Brief Overview of Speech Synthesis and Recognition Technologies*

David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

*Preparation of this report was supported, in part, by Air Force Contract No. AF-F-33615-83-K-0501 with Armstrong Aerospace Medical Research Laboratory (AFSC), Wright-Patterson AFB, OH and, in part, by NIH Research Grant NS-12179 to Indiana University in Bloomington. Portions of this paper were presented at the annual meeting of the Human Factors Society, Dayton, OH, October 3, 1986.

Abstract

An overview of several aspects of speech synthesis and recognition technologies is provided as background for subsequent speakers in this session. Specifically, we discuss speech synthesis by rule using automatic text-to-speech conversion and speaker-dependent isolated word recognition. Both of these speech I/O technologies have been developed sufficiently to the point where commercial products are now available for a number of applications. Some of the limitations of these devices are described and suggestions for future research in both synthesis and recognition are outlined.

A Brief Overview of Speech Synthesis and Recognition Technologies

Speech is without question the most natural means of communication between humans (Lindgren, 1967). It is automatic, requires little conscious effort or attention, and creates few if any demands while other tasks are being carried out concurrently, especially tasks which may require active use of the hands or eyes in demanding conditions. One potential use that speech can serve is as an interface to computers. At the present time, most users interact with computers using traditional screens and keyboards. These systems can and will eventually be replaced by speech I/O. Speech is not only more natural for humans to use, but it is also faster and less prone to errors. While generic speech interfaces to computers are not yet widely available, extensive research efforts have been carried out over the last thirty years to develop speech recognition and synthesis technology so that this goal can be realized.

In this paper, we briefly examine two aspects of this technology, speech synthesis by rule using automatic text-to-speech conversion and speaker-dependent isolated word recognition. Both technologies have been developed sufficiently to the point where commercial products are now available for a number of applications. And both technologies have been studied in our laboratory in two separate research projects. Although these speech I/O products are now available commercially, there are still many questions that need to be answered with regard to how humans will interact with this technology and whether the presumed benefits of speech I/O will outweigh the costs associated with the current technology.

With a text-to-speech system, any computer can generate spoken output from a string of characters and therefore can provide the user with a novel speech display instead of the more traditional VDT screen. In some applications, having a speech display may significantly reduce the user's workload and increase operator efficiency in retrieving information from a computer. In other applications, speech displays may provide entirely new methods for retrieving data and other kinds of information from the computer using standard telephone voice and data channels. As the technology becomes more widely known and as the costs decrease, much wider usage can be anticipated since the benefits may be quite substantial in many applications.

With speech recognition technology, even the limited technology of speaker-dependent isolated word recognition, a user can use spoken language as input to a computer system. Commands, data, and messages can be entered via a microphone or over the telephone line directly into files on the host computer without the need for a CRT display or keyboard entry. Combining both synthesis and recognition technologies can provide advantages in hands-busy and eyes-busy environments where the operator has very severe constraints on his/her information processing capabilities. The implications of speech I/O technology as aids for the handicapped and as additions to business and commercial applications should be fairly obvious.

Voice Output and Text-to-Speech Conversion

A text-to-speech system is a device that automatically converts printed orthographic text into spoken output without human intervention of any kind. This process usually takes place immediately in real-time and will accept any text that can be typed at a computer terminal and converted into ASCII code. Several currently available text-to-speech systems convert unrestricted English text to intelligible speech in real-time. One commercially available system, Infovox, has the capabilities of synthesis of several languages (Magnusson et al., 1984). The speech output generated by a text-to-speech system is synthesized or created anew in real time by the device in response to a phonetic representation of the specific typed input (see Allen, 1973a,b; 1981; Allen, Hunnicutt and Klatt, 1979; Studdert-Kennedy and Cooper, 1966). Most text-to-speech systems are designed to allow the user to customize certain features. For example, there is a "phonetic input mode" that allows the user to specify (with phonetic symbols rather than standard orthography) the correct pronunciation for proper names or to enter a specialized vocabulary that may have unusual pronunciations not captured by the rule system.

One common system of voice output uses stored speech. Natural speech is recorded on audio tape using a microphone and tape recorder. This speech is then digitized with a computer using a analog-to-digital converter. The process involves sampling the speech waveform at a rapid rate and storing the samples in digital form. Typically about 8 to 10 thousand samples are taken for every one second of speech. Unfortunately, for long passages of speech, the storage requirements are enormous. However, there is a good reason to use stored speech. All the digital samples can be retrieved from the computer memory and then reconverted to analog form using a digital-to-analog converter. This process reproduces the speech that was originally recorded with little or no degradation or effects on intelligibility. While there may be some loss in speech quality due to the sampling rate and number of bits used to encode the speech waveform, the resulting speech quality is highly acceptable. These observations are also appropriate for wide-band digitally-encoded speech using a variety of coding algorithms.

When the vocabulary becomes very large and the potential set of messages is theoretically unrestricted, a voice output system using stored speech becomes impractical and extremely expensive (see Cooper, 1963; Cooper, Gaitenby, Mattingly and Umeda, 1969; Studdert-Kennedy and Cooper, 1966). Furthermore, when individual stored items are combined into word strings without additional processing and smoothing, the resulting speech lacks normal pitch and intonation; listeners often describe this type of speech as unnatural and mechanical sounding. The intelligibility of this kind of concatenated speech is often quite poor even though the intelligibility of individual words is typically quite high.

Voice output using stored speech may be contrasted with voice output using various synthesis by rule techniques. In this case, the speech is generated by a series of rules which are used to create utterances on demand (Allen, 1973; Allen et al., 1979; Cooper, 1963). These voice output systems consist of a number of modular subsystems each of which has a special set of rules. The initial typed input is first converted into ASCII code. In most current systems, the ASCII code is then processed through several modules which serve to produce a detailed phonetic description (see Allen, 1981).

In one system, MITalk-79, this analytic process involves the determination of the underlying phonemic, syllabic, morphemic, and syntactic form of the input message as well as adjusting the input when numerals, abbreviations, and special symbols are present. After the basic modules have operated on the input message, any word that has not been analyzed is processed through a set of letter-to-phoneme rules. Once the text has been converted into a phonetic transcription, other modules containing detailed phonological, pitch, stress, and timing adjustments operate on this representation. Additional rules are included to make the speech sound less mechanical. Some rules "smooth" the speech and lead to more natural sounding output. Other rules serve to disambiguate words such as "read" which can be pronounced like "red" or like "reed".

After the input text has been analyzed, it is converted into spoken output. The output process is also modular in nature. Several modules are used to specify the way each speech sound is to be pronounced, how certain speech sounds are modified by specific contexts, and where stress is to be placed. The more detailed the rule system, the closer the synthesized speech approximates natural speech. All the parametric information that has been accumulated in the various modules is then input to a digital speech synthesizer and a speech waveform is generated. Finally, the speech samples are converted to analog form via a digital-to-analog converter and low-pass filtered. The text-to-speech systems that are available at this time all work in real time, performing the analysis and synthesis immediately after the text is input to the device (see Allen, 1981; Bruckert, 1984; Groner, Bernstein, Ingber, Pearlman, and Toal, 1982 for further details).

With error rates for segmental intelligibility of isolated monosyllabic words in the range of 3% to 4% for the best text-to-speech system tested to date, performance is rapidly approaching asymptote (Greene, Logan and Pisoni, 1986; Pisoni, Nusbaum and Greene, 1985). A great deal of further refinement and research probably will be necessary to improve segmental intelligibility much above these levels of performance. At this time, it is probably more productive to look for ways to improve prosody -- the amplitude, timing and durations of individual sounds and words in sentences and the perceived naturalness of synthetic speech. There is a belief among speech researchers that the mechanical sounding quality of synthetic speech is primarily related to the poor knowledge of prosody and the relatively simple algorithms that are currently used to compute pitch and duration in sentences. There is also a need to improve the naturalness of synthetic speech and for further investigations of the factors that control a listener's preference for one synthetic voice over another (Logan and Pisoni, 1986). It is also very likely that the specific application will play an important role in influencing judgments of naturalness and preference among synthetic voices. For the present, however, our studies demonstrate that very high-quality synthetic speech is commercially available and can be incorporated into a wide variety of applications requiring voice output of unrestricted English text.

Speech Recognition

A speech recognition system is a device that takes spoken input (letters, digits, words, or sentences) from a human operator and converts it into some digital representation that can be input to a computer. Compared to the developments in speech synthesis and automatic text-to-speech conversion, the field of speech recognition is still in its infancy (Lea, 1980). Of the commercially available devices, almost all use traditional pattern-matching recognition techniques and require cooperative talkers in relatively benign conditions. Almost all of these systems are speaker-dependent; that is, they require some form of training or "enrollment" to develop "templates" of a given talker's vocabulary. Whenever a new talker uses the system, his/her templates must be accessed from some storage medium or the system must be re-trained with the new talker. While there are some systems that claim to provide speaker-independent recognition, the vocabulary size is typically much smaller and restricted to some well-defined set of words (i.e., digits, or "yes" and "no").

Most commercially available speech recognition systems are, for the most part, only able to recognize discrete utterances -- either isolated words or short phrases. These systems do not "understand" or comprehend the linguistic message in any conventional sense. They operate exclusively on representations of the physical properties of the speech waveform, not on more abstract representations of the content of the message. As a result, they are extremely sensitive to factors that affect the acoustic-phonetics of the speech signal such as noise, stress, fatigue and phonological environment. Moreover, they require that only a single talker use the system at any time. The pattern recognition technology developed for isolated word recognition has been extended to connected speech recognition. Instead of discrete single word utterances, the system is trained on short phrases of connected speech such as a sequence of letters or digits. Most systems do not perform segmentation and treat each utterance as a wholistic pattern regardless of its internal structure.

Continuous speech recognition systems using large vocabularies and unconstrained syntax represent the final objective of speech researchers. With this technology, a talker uses ordinary language input with no constraints on speaking style, vocabulary, or syntax. In some sense, this is like talking to another human. Unfortunately, at present, speech recognition technology is not advanced enough to support speaker-independent continuous speech recognition, an elusive goal that requires substantially more basic speech knowledge and understanding than is currently available. The most serious technical problems in speaker-independent recognition are the enormous variability in the speech waveform produced by the talker, the effects of context, and the surrounding acoustic environment on the talker's articulation of speech sounds in sentences and connected speech. Until these problems are solved, it is doubtful whether much progress will be made in developing robust speech recognition systems that are capable of accepting continuous speech input from a wide variety of talkers using a large vocabulary in an unrestricted semantic domain. It is apparent to some investigators that the solution to this problem will involve the use of novel technologies that rely heavily on acoustic-phonetic knowledge. Traditional pattern recognition techniques are simply not adequate.

Some efforts have been made in the interim to deal with the problems of segmentation in continuous speech recognition by requiring talkers to insert pauses between words and by using very limited vocabularies and constrained syntax in restricted semantic domains such as an office correspondence task. However, a great deal more knowledge of speech production and the acoustic-phonetic properties of speech will be needed before the speech recognition problem can be solved. Traditional pattern recognition techniques are not adequate to overcome the enormous problems of variability in the speech waveform and the multiple sources of knowledge that human listeners routinely use in communicating with each other using spoken language. Moreover, knowledge of language structure will have to be incorporated to resolve ambiguities and deal with impoverished information.

In addition to these technical problems, there are a host of issues that surround the way human talkers interact with speech recognition technology. Talking to a machine is not like talking to another human being who shares a great deal of knowledge and background that is essential for comprehending the message. For some applications, such as a dictating task, humans may have to consciously modify the way they talk by inserting pauses between words and articulating their speech in citation format. For more severe environments where there is noise, psychological stress or high cognitive load, robust recognition algorithms will have to be designed to be self-adaptive (Makhoul, 1985). That is, the algorithms will require "knowledge" of the ways in which talkers modify their speech in noise or under stress or high cognitive load (Pisoni, Bernacki, Nusbaum and Yuchtman, 1985). To accomplish these goals, further basic research will be needed to learn more about the ways human talkers modify their speech output. Finally, substantial efforts will be needed to develop realistic and efficient methods for training and enrollment. The acoustic-phonetic properties of speech change in a variety of ways as a function of time, emotional state, and physiological condition.

Of the many components in a speech recognition system, the human talker is probably the easiest to control or modify. With directed feedback and specialized training procedures, it is possible to change the way a human talks in a relatively short period of time (Zoltan-Ford, 1984). Unfortunately, relatively little research has been directed at this central component of a speech recognition system. Of course, this is not surprising because the bulk of the speech recognition problem has traditionally been conceptualized as an engineering problem rather than a human factors problem (McCauley, 1984). Hopefully, more efforts by human factors specialists will contribute to the development of new and more robust algorithms for speech recognition that can overcome some of the inherent limitations of the traditional pattern recognition techniques that are currently used in the commercial products available today at relatively modest cost (Simpson, McCauley, Roland, Ruth and Williges, 1985).

References

Allen, J. (1973). Speech synthesis from unrestricted text. In J.L. Flanagan and L. R. Rabiner (Eds.), *Speech Synthesis*. Stroudsburg, PA: Dowden, Hutchinson and Ross, pp. 416-428.

Allen, J. (1981). Linguistic-based algorithms offer practical text-to-speech systems. *Speech Technology*, *1*, 1, 12-16.

Allen, J., Hunnicutt, S. and Klatt, D.H. (Eds.). (1979, July). *Conversion of Unrestricted Text to Speech*. Notes for MIT Summer Course 6.69s.

Bruckert, E. (1984). A new text-to-speech product produces dynamic human-quality voice. *Speech Technology*, *2*, 2, 114-119.

Cooper, F. S. (1963). Speech from stored data. *IEEE International Convention Record*, *7*, 137-149.

Cooper, F. S., Gaitenby, J. H., Mattingly, I. G., and Umeda, N. (1969). Reading aids for the blind: A special case of machine-to-man communication. *IEEE Transactions on Audio and Electroacoustics*, *AU-17*, 266-270.

Greene, B. G., Logan, J. S.. and Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, and Computers*, *18*, 100-107.

Groner, G. F., Bernstein, J., Ingber, E., Pearlman, J., and Toal, T. (1982). A real-time text-to-speech converter. *Speech Technology*, *1*, 2, 73-76.

Lea, W. A. (1980). Speech recognition: Past, present, and future. In W. A. Lea (Ed.), *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice-Hall, pp. 39-98.

Lindgren, N. (1967). Speech - man's natural communication. *IEEE Spectrum*, *4*, 75-86.

Logan, J.S. and Pisoni, D.B. (1986). Preference judgements comparing different synthetic voices. *Journal of the Acoustical Society of America*, *79*, S24-S25.

Magnusson, L., Blomberg, M., Carlson, R., Elenius, K., and Granstrom, B. (1984). Swedish speech researchers team-up with electronic venture capitalists. *Speech Technology*, *2*(Jan./Feb.), 15-24.

Makhoul, J. (1985). Automatic speech recognition in severe environments: A report on the National Research Council study. *Proceedings of Speech Tech-85*. NY: Media Dimensions, pp. 228-229.

McCauley, M. E. (1984). Human factors in voice technology. In F. A. Muckler (Ed.), *Human Factors Review: 1984*. Santa Monica, CA: Human Factors Society, pp. 131-166.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C. and Yuchtman, M. (1985). Some acoustic-phonetic correlates of speech produced in noise. *Proceedings of the ICASSP-IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, April, 1985, pp. 1581-1584.

Pisoni, D.B., Nusbaum, H.C. and Greene, B.G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, *73*, 11, 1665-1676.

Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., and Williges, B. H. (1985). System design for speech recognition and generation. *Human Factors*, *27*, 115-141.

Studdert-Kennedy, M. and Cooper, F.S. (1966). High-performance reading machines for the blind. In R. Dufton (Ed.), *Proceedings of the International Conference on Sensory Devices for the Blind*. London: St. Dunstans, pp. 317-342.

Zoltan-Ford, E. (1984). Reducing variability in natural language interactions with computers. *Proceedings of the Human Factors Society 28th Annual Meeting* (Vol. 2). Santa Monica, CA: Human Factors Society.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Developing Methods for Assessing the Performance of Speech Synthesis and Recognition Systems*

David B. Pisoni and Howard C. Nusbaum

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported, in part, by Air Force Contract No. AF-F-33615-83-K-0501 from the Armstrong Aerospace Medical Research Laboratory (AFSC), Wright-Patterson AFB, OH and, in part, by a contract from the IBM Corporation, Yorktown Heights, NY to Indiana University in Bloomington. Portions of this paper were presented at the annual meeting of the Human Factors Society, Dayton, OH, October 3, 1986.

Abstract

As speech I/O technology develops and improves, there is an increased need for standardized methods to systematically assess the performance of these systems. At the present time, speech synthesis and speech recognition technologies are at different levels of maturation and, accordingly, the procedures for testing the performance of these systems are at different stages of development. In the present paper, we describe the results of testing several text-to-speech systems using traditional intelligibility measures. In addition, we outline the design and philosophy of an automated testing procedure for measuring the performance of isolated utterance speaker-dependent speech recognition systems.

Developing Methods for Assessing the Performance of Speech Synthesis and Recognition Systems

Speech synthesis and recognition may be thought of as paradigm cases of emerging technologies. Although currently at different stages of development, both synthesis and recognition are still in their infancy in terms of commerically available products that can be interfaced successfully into the work environment. As is often the case with new technologies, few if any government standards are available while the technology is being developed. Indeed, the standards usually appear after the technology is developed and put in place in some application. Those informal or "unofficial" standards that are available have often evolved more or less informally by researchers and vendors working on developing the technologies.

At the present time, there are no government standards for assessing the performance of speech synthesis and speech recognition systems. The National Bureau of Standards (NBS) has been working on these problems for a number of years and has issued guidelines for performance testing of speech recognition systems (Pallett, 1985); however, no formal standards are presently available for either synthesis or recognition technology. In the absence of any formal standards, we have carried out a number of studies to assess and compare the performance of synthesis and recognition devices. Over the last seven years, efforts have been made in our laboratory to collect relevant behavioral data from several synthesis by rule systems and to draw comparisons among several commerically available text-to-speech systems. In some sense, our studies on this problem over the last seven years have resulted in a "de facto" standard for assessing the segmental intelligibility of synthetic speech produced by rule. We have collected enough data from a variety of systems under the same set of conditions to permit meaningful comparisons among several systems. Although our tests were conducted under laboratory conditions and some caution must be exercised in generalizing to other environments, our data have served as an extremely useful benchmark and have laid the foundation for the development of other behavioral tests to assess the performance of speech synthesis systems.

Most researchers will probably argue that these behavioral tests are far from the ideal instruments to assess performance of speech synthesis systems. However, the tests we have used have provided extremely reliable data and they are able to discriminate among the various systems tested in a meaningful way. Those systems that sound good to a naive listener in informal listening tests are the systems that display high scores on tests such as the modified rhyme test (MRT), a test designed to measure the segmental intelligibility of phonemes in initial and final position of isolated CVC English words. Those systems that do not sound good are ones that show low scores on these tests. In all cases involving evaluation of the intelligibility and/or quality of a speech synthesis system, the human listener is the ultimate discriminator. Consequently, behavioral data are needed in drawing comparisons among systems or ranking the performance of systems on some absolute scale.

The situation with regard to the evaluation of speech recognition technology is somewhat different from the evaluation of speech synthesis technology. First, the state of the art for each technology differs rather substantially at the present time. Speech synthesis technology is currently at the point where any English text can be automatically converted into fluent speech that is only slightly less intelligible than natural speech. However, speech recognition technology is in a much less advanced state because it is not possible for a human to talk to a computer with the same fluency and ease as it is to talk to another human. Second, while we have carried out an extensive program of research investigating a wide range of factors that can influence the performance of a text-to-speech system, to date there has not been the same kind of effort directed at systematically studying the performance of speech recognition systems. There are a large number of variables that can and do affect the performance of a speech recognizer and the contribution of these variables is poorly understood at this time.

In the two sections below we briefly summarize work we have carried out in developing techniques for the evaluation of synthesis and recognition systems. The section on synthesis reports behavioral data obtained from human listeners in a variety of perceptual tests; the section on recognition describes our recent work in developing automated testing procedures to assess the performance of several commercially available speaker-dependent isolated word recognition systems using a large data base of speech collected at Texas Instruments.

Evaluation of Speech Synthesis Systems

To interpret the results of evaluation studies, it is necessary to consider the following underlying factors that affect an observer's performance: (1) the specific demands imposed by a particular task, (2) the inherent limitations of the human information processing system, (3) the experience and training of the human listener, (4) the linguistic structure of the message set, and (5) the structure and quality of the speech signal.

Task Complexity

The first factor that constrains performance concerns the complexity of the tasks that engage an observer during the perception of speech. In some tasks, the response demands are relatively simple, such as deciding which of two known words was said. Other tasks are extremely complex, such as trying to recognize an unknown utterance from a virtually unlimited number of response alternatives while engaging in an activity that already requires attention. There is a substantial amount of research in the cognitive psychology and human factors literature demonstrating the powerful effects of perceptual set, instructions, subjective expectancies, cognitive load, and response set on performance in a variety of perceptual and cognitive tasks. The amount of context and the degree of uncertainty in the task also affect an observer's performance in substantial ways.

Limitations on the Observer

The second factor influencing recognition of synthetic speech concerns the substantial limitations on the human information processing system's ability to perceive, encode, store, and retrieve information. Because the nervous system cannot maintain all aspects of sensory stimulation (and therefore must integrate acoustic energy over time), very severe processing limitations have been found in the capacity to encode and store raw sensory data in the human memory system. To overcome these capacity limitations, the listener must rapidly transform sensory input into more abstract neural codes for more stable storage in memory and subsequent processing operations. The bulk of the research on cognitive processes over the last 25 years has identified human short-term memory (STM) as a major limitation on processing sensory input. The amount of information that can be processed in and out of STM is severely limited by the listener's attentional state, past experience, and the quality of the sensory input.

Experience and Training

The third factor concerns the ability of human observers to quickly learn effective cognitive and perceptual strategies to improve performance in almost any sort of task. When given appropriate feedback and training, subjects can learn to classify novel stimuli, remember complex pattern sequences, and respond to rapidly changing stimulus patterns in different sensory modalities. Clearly, the flexibility of subjects in adapting to the specific demands of a task is an important constraint that must be evaluated, or at least controlled in any attempt to evaluate synthetic speech.

Message Set

The fourth factor relates to the structure of the message set; that is, the constraints on the number of possible messages and the organization and linguistic properties of the message set. This linguistic constraint depends on the listener's knowledge of language.

Signal Characteristics

The fifth factor deals with the acoustic-phonetic and prosodic structure of a synthetic utterance. This constraint refers to the veridicality of the acoustic properties of the synthetic speech signal compared to naturally produced speech.

Speech signals may be thought of as the physical consequence of a complex and hierarchically organized system of linguistic rules that map sounds onto meanings and meanings back onto sounds. At the lowest level in the system, the distinctive properties of the speech signal are constrained in substantial ways by vocal tract acoustics and articulation. The choice and arrangement of speech sounds into words is constrained by the phonological rules of language; the arrangement of words in sentences is constrained by syntax; and finally, the meanings of individual words and the overall meanings of sentences in a text is constrained by semantics and pragmatics. The contribution of these various levels of linguistic structure to perception will vary substantially from isolated words, to sentences, to passages of fluent continuous speech. In addition to linguistic structure, the ambient noise level and the spectro-temporal properties of noise in the environment in which the speech signal occurs will also affect recognition.

Perceptual Evaluation of Synthetic Speech

There are basically three areas in which a text-to-speech system could be deficient that would impact the overall intelligibility of the speech: (1) the spelling-to-sound rules, (2) the computation and production of suprasegmental information, and (3) the phonetic implementation rules that convert the internal representation of phonemes and/or allophones into a speech waveform. In our previous research, we have found that phonetic implementation rules are a major factor in determining the segmental intelligibility of a voice response system (Nusbaum & Pisoni, 1982). The task that is generally used as a standard measure of the segmental intelligibility of speech is the Modified Rhyme Test (MRT). In this procedure, subjects are asked to identify a single word by choosing one of six alternative response words differing by a single phoneme in either initial or final position (House, Williams, Hecker, & Kryter, 1965). All the stimuli in the MRT are CVC words; on half the trials, the responses share the VC of the stimulus and on the other half, the responses share the CV. Thus, the MRT provides a measure of how well listeners can identify either the initial or final phoneme of a set of spoken words. To date, we have evaluated natural speech and speech produced by a number of different text-to-speech systems including the Votrax Type-'n-Talk, the Speech Plus Prose-2000, the MITalk-79 research system, and DECtalk (Greene, Logan, & Pisoni, 1986; Greene, Manous, & Pisoni, 1984). Word identification performance for natural speech was the best at 99.4% correct. For DECtalk, we evaluated speech produced by Paul and Betty, two of DECtalk's nine voices, and found different levels of performance on these voices -- 96.7% of the words spoken by the Paul voice were identified correctly while only 94.4% of Betty's words were identified correctly. The level of performance for the Paul voice comes quite close to natural speech and is higher than performance for any other text-to-speech system we have studied to date. Performance on MITalk-produced speech was somewhat lower than either of the DECtalk voices at 93.1% correct word identification. The early prototype of the Prose-2000 produced speech that was identified at 87.6% correct, although the current Prose-2000 Version 3.0 is considerably improved, with performance at 94.3% correct. Finally, the least intelligible synthetic speech was produced by the Votrax Type-'n-Talk at 67.2% correct word identification. These results, obtained under closely matched testing conditions, show a wide range of variation among text-to-speech systems that seems to reflect the amount of basic research that was carried out to develop the phonetic implementation rules of these different voice response systems.

In addition to these tasks, we have used an open-response format version of the MRT in which listeners are instructed simply to write the word that was heard on each trial. This open-response format provides a measure of performance when constraints on the response set are minimized (compared to the six-alternative forced choice version). It also provides information about the intelligibility of the vowels that is not available in the closed-response set version of the MRT. A comparison of the closed- and open-response versions of the MRT for speech produced by different text-to-speech systems with natural speech indicates the degree to which listeners rely on response-set constraints. Performance on the open-response set MRT for natural speech was at 97.2% correct exact word identification compared to 99.4% correct in the closed-response set task. Even when there are no strong constraints on the number of alternative responses for natural speech, performance is better than for any text-to-speech system with a constrained set of responses. For the MITalk-79 research system, performance in the open-set task is considerably worse than at 75.4% correct. Similarly, DECtalk's Paul voice produced words that were identified at the 86.7% level. These results show a large and reliable interaction between intelligibility measured in the closed-response format MRT and the open-response format MRT. Even though the rank ordering of intelligibility stays the same across the two forms of the MRT, it is clear that as speech becomes less intelligible, listeners rely more heavily on response-set constraints to aid recognition.

Evaluation of Speech Recognition Systems

Measuring the performance of a speech recognition system involves consideration of several closely related factors that interact to influence the final observed performance of the system as a whole. Several human factors problems can be identified in characteristics of the user and the associated individual differences and variability in talking style. Other problems are related to the design of the user/system interface and the applications environment in which the recognizer will be used. The specific task plays an important role in affecting performance depending on whether the system is used for dictation, quality control inspection, command and control, or voice data entry using a small vocabulary. Finally, characteristics of the recognition system itself, the signal processing algorithms and overall system architecture, also play an important role in recognition performance. If the recognition algorithms are inherently incapable of discriminating fine phonetic details, the system will not be able to recognize phonetically confusable vocabularies such as the alphabet. Similarly, properties of the physical environment such as the microphone and the surrounding ambient noise level also affect performance of the system.

Vocabulary

One major factor that influences performance of a speech recognizer is the vocabulary and the inherent acoustic-phonetic similarity of members of the ensemble of to-be-recognized utterances. The letters of the alphabet are notoriously difficult to recognize whereas words like "Presbyterian," "Episcopalian," and "chrysanthemum" are extremely easy. The differences in performance between these two vocabularies are intimately related to the number of phonetically similar words in the vocabulary and the confusability of the sound patterns. As the length of a word in phonemes increases, it becomes more and more unique and distinctive from other phonetically similar words. In the case of the alphabet, particularly the so-called "E-set" (b,p,d,t,c,z,e), the items represent minimal pairs of sound contrasts that are difficult to discriminate in isolation even by human observers who are often considered the benchmark against which performance of a speech recognizer is ultimately evaluated. The differences among members of the "E-set" are restricted to the beginning of the utterances and there is little additional acoustic-phonetic redundancy that can be used to discriminate between them.

Talker Variability

Another factor that affects recognizer performance involves the problem of talker variability. This problem is so important and so central to solving the major problems in speech recognition that a distinction is drawn in describing speech recognizers between speaker-dependent and speaker-independent recognition systems. The former represent systems that are designed to recognize utterances from a specific individual talker; the latter describe systems that can recognize utterances from any talker. Speaker-dependent systems require some period of training and enrollment so that the vocabulary of the talker can be entered into the system and a set of templates can be developed for that particular talker. Speaker-independent systems can operate without a training period and are, in principle, capable of recognizing the speech of any talker presented to the system.

The problem of dealing with talker variablity is therefore reduced rather substantially with speaker-dependent recognition systems although it is not completely eliminated by simply restricting the utterances to a single talker who has trained the system at some earlier time. Restricting the vocabulary to a single talker does not eliminate within-talker variability due to stress, fatigue, emotional state and the momentary changes in cognitive load. Moreover, environmental conditions play a significant role in affecting recognizer performance even with speaker-dependent systems. Substantial decrements in recognition performance are routinely obtained in the presence of environmental noise or when the utterances are transmitted over conventional telephone lines compared to benign laboratory conditions. Another important related issue in speech recognition deals with how performance in the laboratory will generalize to performance in an application-specific context in which a variety of uncontrolled environmental factors are operating. Relatively little systematic research has been done on this problem.

Recently, we began a research project that was designed to carry out systematic and controlled laboratory benchmark tests of the performance of isolated-utterance, speaker-dependent speech recognition systems. Performance testing for a recognition system typically requires training the recognizer on several tokens of a vocabulary and then carrying out recognition tests on a number of other tokens of the vocabulary. This training and testing protocol must be carried out for a number of talkers and some tests may require as many as 5120 trials to collect sufficient recognition data to assess performance. In the past, these tests were carried out trial by trial by a human operator requiring a great deal of time and effort.

Automated Recognition Testing

To reduce the costs and effort involved in carrying out performance tests and to reduce the possibility of operator error in testing, we have developed a computer-controlled testing system that permits a researcher to define and then execute automatically training and testing protocols for different recognition systems. The basic concept for this testing system derives from our research in testing human subjects in speech perception experiments. In these experiments, human subjects are tested under real-time control of a computer that presents speech signals over headphones and collects the subjects' responses to these stimuli. Speech signals are stored in digital waveform files on a large disk and are retrieved on demand and converted to analog form for presentation to subjects. In our testing procedure, the subject is a particular speech recognition system interfaced to a microcomputer. A different computer controls the training and testing protocols and presents speech signals to the recognizer. The recognizer responds to the controller through its host microcomputer over a serial communications line.

We have implemented this testing paradigm by dividing the test control system into two parts: (1) a virtual device controller and interface (VDC) for generalized speech recognition systems and (2) a device dependent interface (DDI) that is specifically programmed for each individual recognizer. The VDC embodies a model of a generic speech recognition system. This model includes several functions and parameters that can be manipulated by an operator. The operator uses a command language to program the generic recognition model in the controller. The VDC communicates these commands to the DDI which translates the generic recognizer commands into the specific commands and syntax of the recognizer that is being tested. Thus, the DDI serves as the communications host and translator for the recognition system that is being tested.

We are currently testing recognition systems using the digital speech database collected by Texas Instruments (Doddington & Schalk, 1981). This database consists of two vocabularies: (1) the TI-20 consisting of the ten digits and ten control words, and (2) the 26 letters of the alphabet. The database consists of ten training tokens and sixteen testing tokens of each vocabulary item. Eight male and eight female talkers produced the speech which was digitized at 12.5 kHz with 16 bits of resolution.

Prior to testing, signal levels presented to the recognition system are adjusted to the optimal level for a specific recognizer. Following calibration, each recognizer is trained and tested on one talker at a time, for each talker in the database. Recognizers are only tested on one vocabulary at a time (either the TI-20 or the alphabet). A minimum of three tests is carried out for each recognizer on each vocabulary. In each test, a recognition system is trained on a different number of tokens from the training set of the speech database (either one, two, or three tokens of each word). A comparison across these tests for each recognizer indicates the incremental improvement in recognition performance as a function of increased training. The performance of different recognition systems is compared across training curves with a specific vocabulary to determine the relative performance of each system. Beyond these basic three tests, other tests have been carried out to try to improve the recognition performance of a specific system to its optimal level and to determine the effects of noise on recognition performance.

Although it is clear that these tests cannot, by themselves, completely assess the performance of currently available speech recognition systems, they do provide an index of the relative performance of these devices under controlled laboratory testing conditions that will permit objective comparisons of different systems. These data are therefore a much better measure of recognition performance than the performance figures typically cited by vendors for their systems which are collected under uncontrolled and unspecified testing conditions. Furthermore, although it is an emprical question that is yet be resolved, it is entirely possible that the rank ordering of performance on laboratory benchmark tests may accurately predict the rank ordering of performance under application-specific conditions. From our own work on this problem, it is clear that a great deal more research is needed using systematic testing of recognition systems under a wide range of conditions and applications. By comparison with the performance testing of text-to-speech systems, systematic laboratory-based performance testing of speech recognition systems has only just begun.

References

Doddington, G. R., & Schalk, T. B. (1981). Speech recognition: Turning theory into practice. Spectrum, September, 26-32.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments, & Computers, 18, 100-107.

Greene, B. G., Manous, L. M., & Pisoni, D. B. (1984). Perceptual evaluation of DECtalk: A final report on Version 1.8. Research on Speech Perception Progress Report No. 10, Bloomington, IN: Speech Research Laboratory, Indiana University.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. (1965). Articulation testing methods: Consonant differentiation with a closed-response set. Journal of the Acoustical Society of America, 37, 158-166.

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. Behavior Research Methods, Instruments, & Computers, 17, 235-242.

Pallett, D. S. (1985). Performance assessment of automatic speech recognizers. Journal of Research of the National Bureau of Standards, 90, 371-387.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Recognition Performance of Six Isolated Utterance Speech Recognition Systems*

Howard C. Nusbaum, C. Noah Davis, David B. Pisoni, and Ella Davis

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This research was supported by a contract from IBM Corporation to Indiana University in Bloomington. We thank Texas Instruments, Interstate Electronics, and Votan for providing speech recognition devices and assistance during testing. We also thank Dr. David Pallett at the National Bureau of Standards for providing the speech databases that we used in performance testing. We thank Dr. Moshe Yuchtman for his invaluable assistance in collecting data. Special thanks are also extended to Jerry Forshee and David Link for their help with instrumentation. This report is the written version of a paper presented at the 1986 Voice Data Entry Systems Applications Conference, Alexandria, VA.

Abstract

Although performance data are often freely cited by vendors of speech recognition devices, the conditions under which these data were collected are seldom specified in detail. Thus, it is nearly impossible to interpret performance comparisons among different recognition systems. To directly compare performance of six commercially available speech recognition systems, we developed a computer-controlled testing system and a set of standard tests. We carried out these tests to assess the performance of recognition devices sold by Texas Instruments, Votan, Dragon Systems, IBM, Interstate Electronics, and NEC. Our results demonstrated several reliable performance differences among these systems. However, in general, performance differences among these devices were quite small and were reduced by appropriate training. The results also indicated that the effects of training on performance are much more pronounced for difficult vocabularies, such as the alphabet than easy vocabularies, such as the digits. Finally, the results for recognition of the speech of one talker in the testing database suggest that user-specific difficulties in recognition performance may, in some cases, result from an interaction among the application vocabulary, the user's speech, and the training algorithm used in the recognition device.

Recognition Performance of Six Isolated Utterance Speech Recognition Systems

Over two years ago, at the Speech Research Laboratory at Indiana University, we began a research project directed at measuring and comparing the performance of commercially available, speech recognition systems. As experimental psychologists, we decided to test the performance of these systems using the same general approach that we have taken in testing the perception of speech signals by human listeners. In this paradigm, we present human listeners with digitized speech signals and collect their responses under computer control. Thus, we had three basic objectives in carrying out this research.

The first goal of this project was to develop an automated system that would allow us to carry out performance tests with speech recognizers completely under computer control. This system was designed to take a program as input that describes the entire training and testing protocol for one recognition system and collect performance data without any human intervention whatsoever. The second objective of this research was to develop a methodology for measuring the performance of speaker-dependent, isolated-utterance, speech recognition systems. These testing procedures were realized as a set of programs that are interpreted by the automatic testing system. The final goal was to actually test several commercially available speech recognition systems and compare their performance using our testing methods.

Methods

Recognition Devices. Table 1 shows the speech recognition systems that were tested along with the approximate date when testing began for each system. Vocabulary sizes in these systems range from 50 to 256 utterances. All systems were tested as isolated-utterance recognition systems, even though some of them used connected-speech recognition algorithms. In order to facilitate the testing process, only IBM-PC compatible recognition systems were used. This constraint was imposed by our automatic testing system which uses an IBM-PC as a host for the recognition device.

Insert Table 1 about here

Speech Databases. All recognition devices were tested using two digitized speech databases. One database, called the TI-20, consisted of spoken digits and ten control words such as GO, ENTER, RUBOUT, and REPEAT. The second database, the TI-Alphabet, consisted of the spoken letters of the alphabet. Although the TI-20 provides a more realistic application vocabulary, we included both databases in our tests for two reasons. First, the TI-20 is a relatively easy vocabulary, containing few confusable words. Second, this database has been in the public domain for some time, and has been used as the basis for developing and improving recognition algorithms by a number of vendors of speech recognition technology. By comparison, the

Table 1

Recognition Devices and Test Dates*

System	Test Date
NEC SR-100	6/85
Votan VPC-2000	9/85
Interstate Vocalink	10/85
Dragon Systems	7/86
TI Speech-II	7/86
IBM System	9/86

---

*Note: All speech recognition systems are small vocabulary (<300 items), speaker-dependent, IBM-PC compatible devices costing less than $3,500.

TI-Alphabet provides a more difficult vocabulary because of the highly confusable members of the E-set -- the letters B, D, G, C, P, T, E, Z, and V. Also, this database has not generally been used as the basis for algorithm development. Both databases therefore provided a means of investigating a range of recognition performance with each device we studied.

Procedure. Prior to testing each recognition device, the output level of a preamplifier was set based on the outcome of two types of calibration tests. One series of tests consisted of a complete recognition test using the TI-20 database at signal levels varying from 2.5 to 12.5 dB in 2.5 dB steps. In the second set of tests, we trained and tested each recognizer on the same set of digits to find the level that produced the smallest distance scores. One signal level was chosen that produced the fewest errors in the first set of tests and the smallest distance scores in the second series of tests.

Each recognition device was tested at the chosen signal level on the TI-20 database and on the TI-Alphabet database. A minimum of three tests was carried out on each database to investigate the effects of the number of training tokens on performance. A test was carried out with one training token, three training tokens, and five training tokens. Very few of the recognizers were capable of using more than five training tokens. However, we have carried out tests in which recognizers such as the Dragon system were trained with all ten tokens. It is important to note that the NEC SR-100 does not update its template representations when trained on more than one item. To simulate updating on the NEC system, we created one vocabulary item for each training token. During recognition, we mapped the set of tokens corresponding to one word onto a single response.

In all tests, the reject threshold was turned off to elicit only substitution errors. A substitution error occurs when a recognition system responds to an utterance with an incorrect vocabulary item.

Description of SPERTES

Our computer-controlled testing system is called SPERTES (SPEech Recognition TEsting System). The system consists of two major components. A Vax-11/750 controls each testing session by interpreting a script of generic recognition commands. An IBM-PC serves as the host and interface for the recognition device. Training and testing is carried out under the direction of a Virtual Device Controller (VDC) running on the Vax. This program embodies a generic model of a recognition system and interprets scripts that describe training and recognition in terms of generic commands. These commands are sent to a Device Dependent Interface (DDI) on the PC which translates commands into recognizer-specific format. The DDI also returns responses from the recognizer recoded into a generic format.

Insert Figure 1 about here

Figure 1 shows the major components of SPERTES. The Vax and PC communicate over a serial line. The digital speech database is stored on large-capacity, high-speed disks on the Vax. Digitized speech is converted to analog form by the DSC and is then presented directly to the recognition

SPERTES: Speech Recognition Testing System

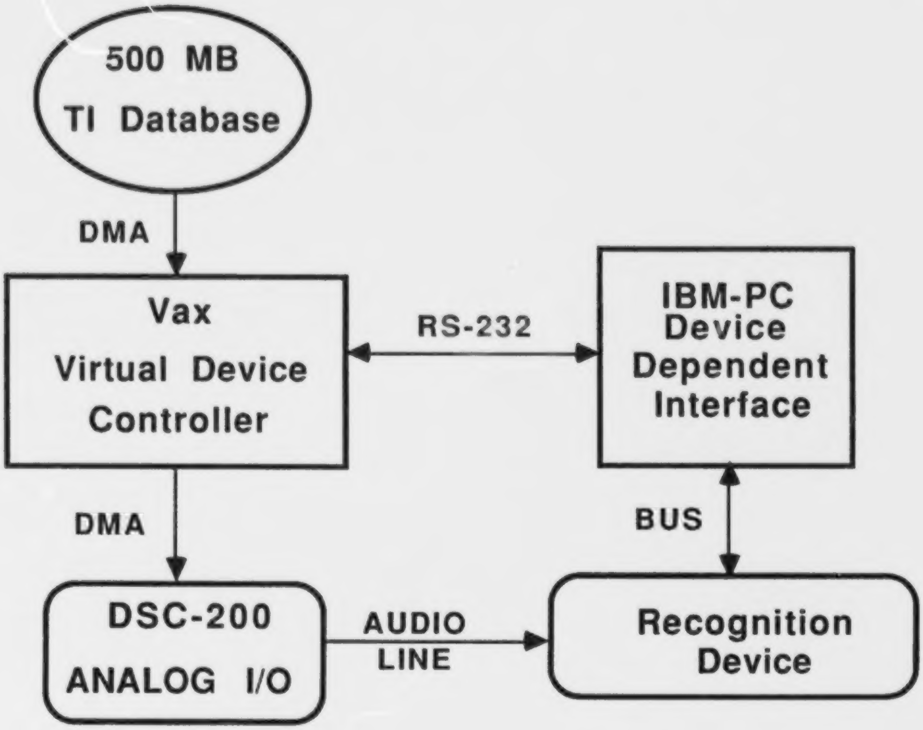

500 MB
TI Database
DMA
Vax
Virtual Device
Controller
RS-232
IBM-PC
Device
Dependent
Interface
DMA
BUS
DSC-200
ANALOG I/O
AUDIO
LINE
Recognition
Device

Figure 1. SPERTES consists of a VAX-11/750 that serves as a Virtual Device Controller (VDC) and an IBM-PC that serves as a Device Dependent Interface (DDI). The VDC controls the training and testing protocol for a recognition device that is interfaced to the DDI.

device under control of the Vax. Each recognition device is directly interfaced to the PC which serves as a host.

Results and Discussion

The results of testing on the TI-20 vocabulary are shown in the top panel of Figure 2. The mean percentage of substitution errors (averaged across talkers) out of 5120 test tokens is displayed for each of the six recognition devices. Several results are of interest in this graph. First, a general improvement in recognition performance can be observed as the amount of training increased. However, the largest and most reliable effects of training are produced by the first three training tokens.

Insert Figure 2 about here

Another finding of interest is that, although reliable differences in performance are observed among recognition devices trained on one token, these differences are substantially reduced when the systems are trained on five tokens. With one training token, recognition devices can be assigned to three categories based on performance, from lowest to highest accuracy: (1) the NEC and Interstate systems, (2) the IBM and Dragon systems, and (3) the TI and Votan systems. However, with five tokens of training, the Votan, IBM, TI and Dragon systems are nearly equivalent in performance, while the Interstate and NEC systems are less accurate for this vocabulary.

The bottom panel of Figure 2 shows recognition performance for the six recognition systems on the TI-Alphabet database (containing 6656 test tokens). It should be noted that the scales showing performance in percentage of substitution errors are considerably different for the TI-Alphabet and TI-20 databases. Performance is substantially worse for the Alphabet vocabulary due to the presence of sets of confusable letters such as the E-set (e.g., B, D, G, etc.) and the A-set (A, K, J).

Also, by comparison to performance on the TI-20 vocabulary for the Alphabet, all systems showed statistically reliable improvements in recognition performance as the number of training tokens was increased from one to three to five. However, the results for the Votan system only show this improvement when the data for one talker is omitted from the analysis. This talker's data will be discussed separately later.

The bottom panel of Figure 2 also shows that performance differences among recognition devices are substantially reduced with increased training. At five tokens of training, no overall statistically reliable differences in performance were observed among any of the recognition systems. However, with minimal training using only a single token, performance of the Votan and TI systems is reliably better than the other systems.

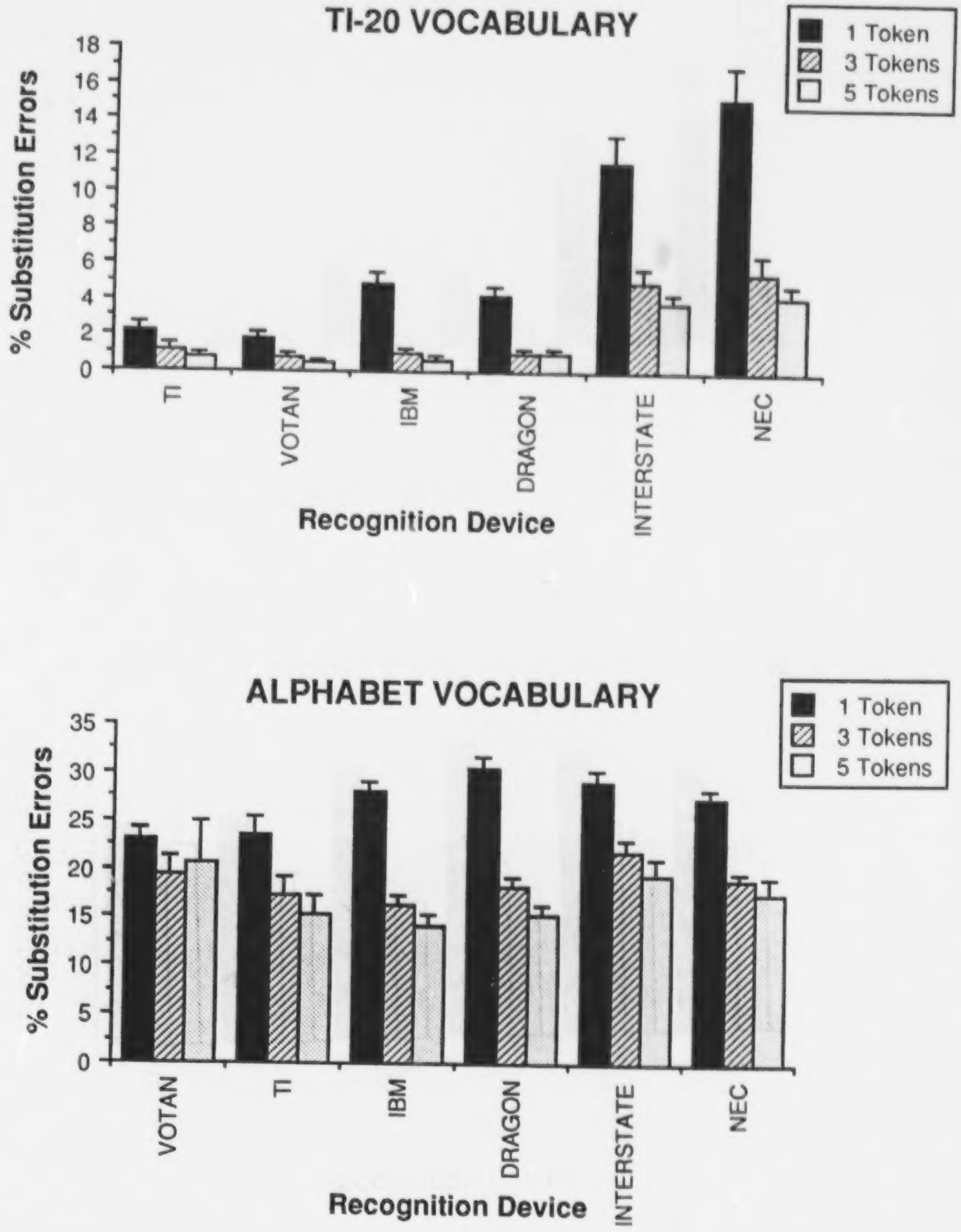


Figure 2. The effect of number of training tokens on recognition accuracy for six commercially available speaker-dependent, speech recognition devices. Recognition performance was tested on the TI-20 vocabulary (top panel) and the Alphabet vocabulary (bottom panel).

Insert Figure 3 about here

The top panel of Figure 3 shows the results for the IBM and Dragon systems tested on the TI-20 database when these devices have been trained on one, three, five, and ten tokens selected from the training items. In addition, these data show the effects of training both the Dragon and IBM systems with ten tokens and with adaptation. The adaptation feature of the Dragon and IBM systems allows the recognition algorithms to update a model of a word throughout recognition, even after explicit training is complete. Although this feature is really intended to improve performance with multiple talkers, or under less than optimal environmental conditions, it also allows the recognizer to use more speech data for improving performance. However, for the TI-20 database, the results indicated no reliable improvements in recognition performance with either ten tokens of training or with adaptation in comparison with either three or five tokens of training. Thus, for the TI-20 vocabulary, training with three tokens yields optimal levels of performance for both the IBM and Dragon systems, despite the ability of these devices to make use of more speech data in training.

Recognition performance on the TI-Alphabet for the IBM and Dragon systems is shown in the bottom panel of Figure 3 for training with one, three, five, and ten tokens. Testing with adaptation was not carried out for the TI-Alphabet with the IBM and Dragon systems. The performance data for the TI-Alphabet show that training with ten tokens produces significantly better performance than the other training conditions. These results indicate that increased training is more important and beneficial to recognition performance with difficult vocabularies, that is, with vocabularies that contain more confusable tokens.

Comparing performance across the TI-20 and TI-Alphabet vocabularies, the largest effect of training is generally provided by the first three tokens. For vocabularies that do not contain many confusable items, recognition performance does not appear to benefit from greater amounts of training, even for systems that use sophisticated Hidden Markov Modeling (HMM) techniques. (HMM techniques permit the use of much more training data than more traditional dynamic time warping, template matching systems.) Furthermore, when recognition systems are trained appropriately (e.g., about 3-5 tokens as per vendors' instructions), equivalent amounts of training result in roughly equivalent performance. However, for vocabularies that contain highly confusable items, increased training beyond three tokens does reliably improve performance. And, HMM-based systems, by virtue of their ability to train on more speech data than other systems, may perform better when trained more extensively.

Goats and Sheep. As noted earlier, unlike other recognition devices, the Votan system did not display reliable increases in recognition performance for the TI-Alphabet vocabulary due to performance of one talker. The top panel of Figure 4 shows the data for this talker -- M1 -- and the data for F1, a more typical talker, for the Votan and Interstate systems on the Alphabet vocabulary. The bottom panel of Figure 4 shows the performance of these two recognition devices for the same talkers on the TI-20 vocabulary.

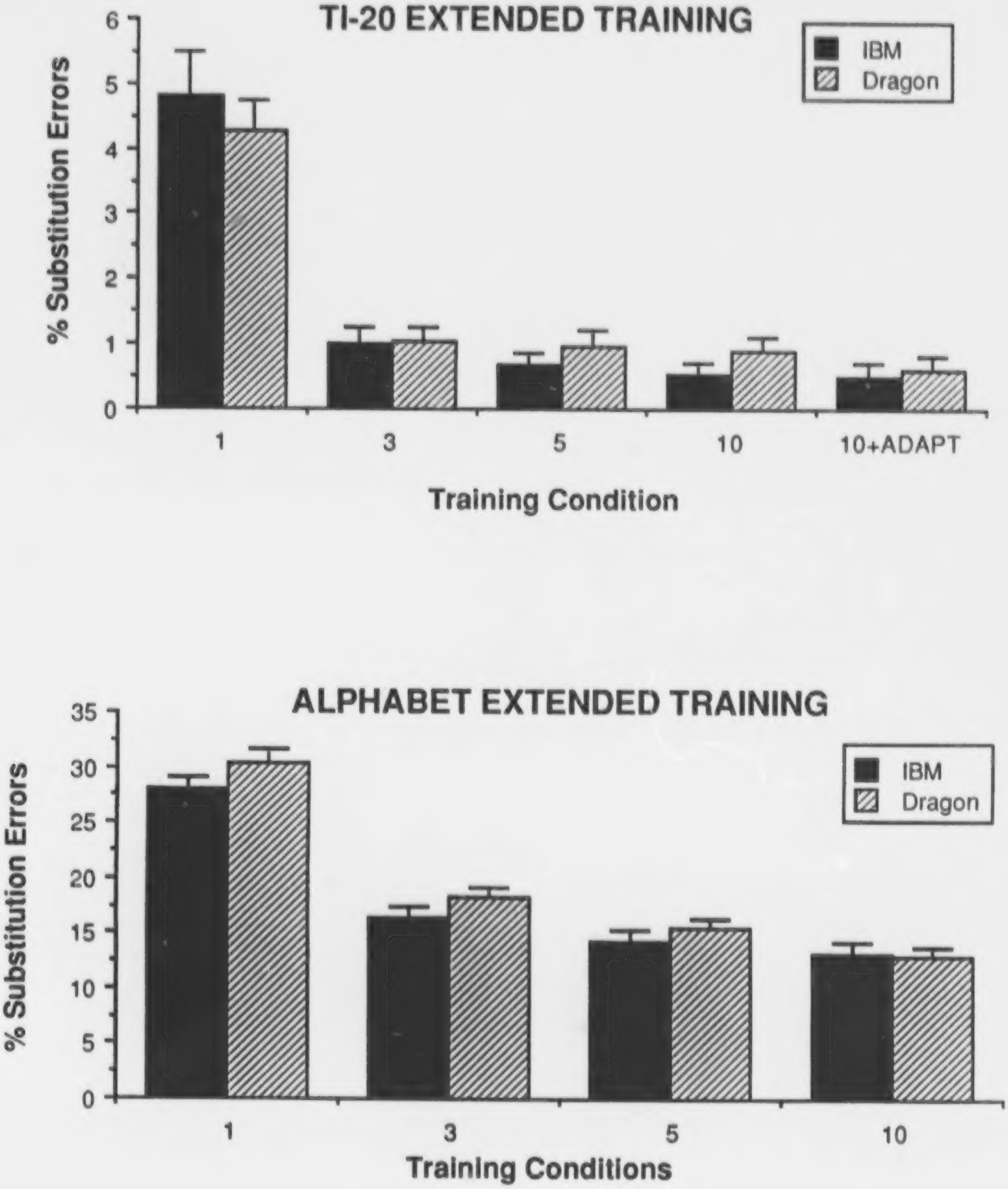


Figure 3. The top panel shows the effects of extended training with 10 tokens and 10 tokens with *adaptation* compared to training with 1, 3 and 5 tokens for the IBM and Dragon speech recognition systems tested on the TI-20 vocabulary. (Note: Adaptation allows the recognition device to modify utterance models throughout recognition.) The bottom panel shows the effects of training with 1, 3, 5, and 10 tokens of training for the IBM and Dragon systems tested on the Alphabet vocabulary.

Insert Figure 4 about here

For both recognition systems tested on the TI-Alphabet and TI-20, performance on F1's speech generally improves as the number of training tokens is increased. A similar pattern can be seen in the performance of the Interstate system on M1's speech for the Alphabet and for the Votan and Interstate systems on the TI-20. However, for the Alphabet, performance of the Votan system on M1's speech decreases with more training, producing an error rate of almost 85% with five tokens of training. Comparing the performance of the Votan system on the Alphabet vocabulary for talkers M1 and F1 might lead one to conclude that M1 is a "goat" -- an unsuccessful user of speech technology, while F1 is a "sheep" -- a successful user of this technology. This conclusion receives further support because performance for one talker shows an error rate that is as high as the recognition accuracy obtained for another talker.

Note that this huge error rate drops back down to about the same performance level as the Interstate for the point on the figure labeled "INDEXED." In the training condition labeled INDEXED, each recognition device was trained on five tokens of speech, but each token was used to create a separately indexed vocabulary item. Using software built into the DDI, we mapped the separately indexed items onto the appropriate word. Thus, during training, each letter of the vocabulary would have five separate templates such as A1, A2, A3, A4, and A5 corresponding to each of the different training tokens. During recognition, if one of these five items was recognized, the DDI would simply respond with the word "A", ignoring which token generated the response. This allowed us to circumvent the mechanism by which a recognition device combines different tokens of speech into a single representation. The data for this INDEXED condition for the Votan demonstrates that the increasingly poor performance with increased training on M1's speech is not due specifically to the recognition algorithm, but instead is a consequence of the method by which the Votan combines tokens of speech during training. It is interesting to note that the increased error rate of this type with training did not occur for any other talker in the database or for the TI-20 vocabulary.

We believe data such as these bear directly on the issue of why differences between goats and sheep are often reported in the literature and discussed frequently among some researchers working in the field. Several hypotheses have been proposed to account for the differences observed in recognition performance by so-called goats and sheep. One hypothesis is that goats and sheep differ in motivation towards using the equipment. A second hypothesis is that goats and sheep differ in the reliability of their speech. Finally, a third hypothesis is that goats and sheep differ in the acoustic-phonetic "clarity" or distinctiveness of their speech. In all three cases, these hypotheses attribute the goats/sheep distinction to the talker alone. However, the results obtained for talker M1 in the present study suggest that in some cases, a goat for one vocabulary or recognition device may be a sheep for a different vocabulary or recognizer. As a consequence, it appears that poor recognition performance for a particular talker may result from an interaction between vocabulary, recognition device, and talker and, therefore, may not be entirely due to the talker alone. We believe this is an important observation concerning the combined effects of all three sources of

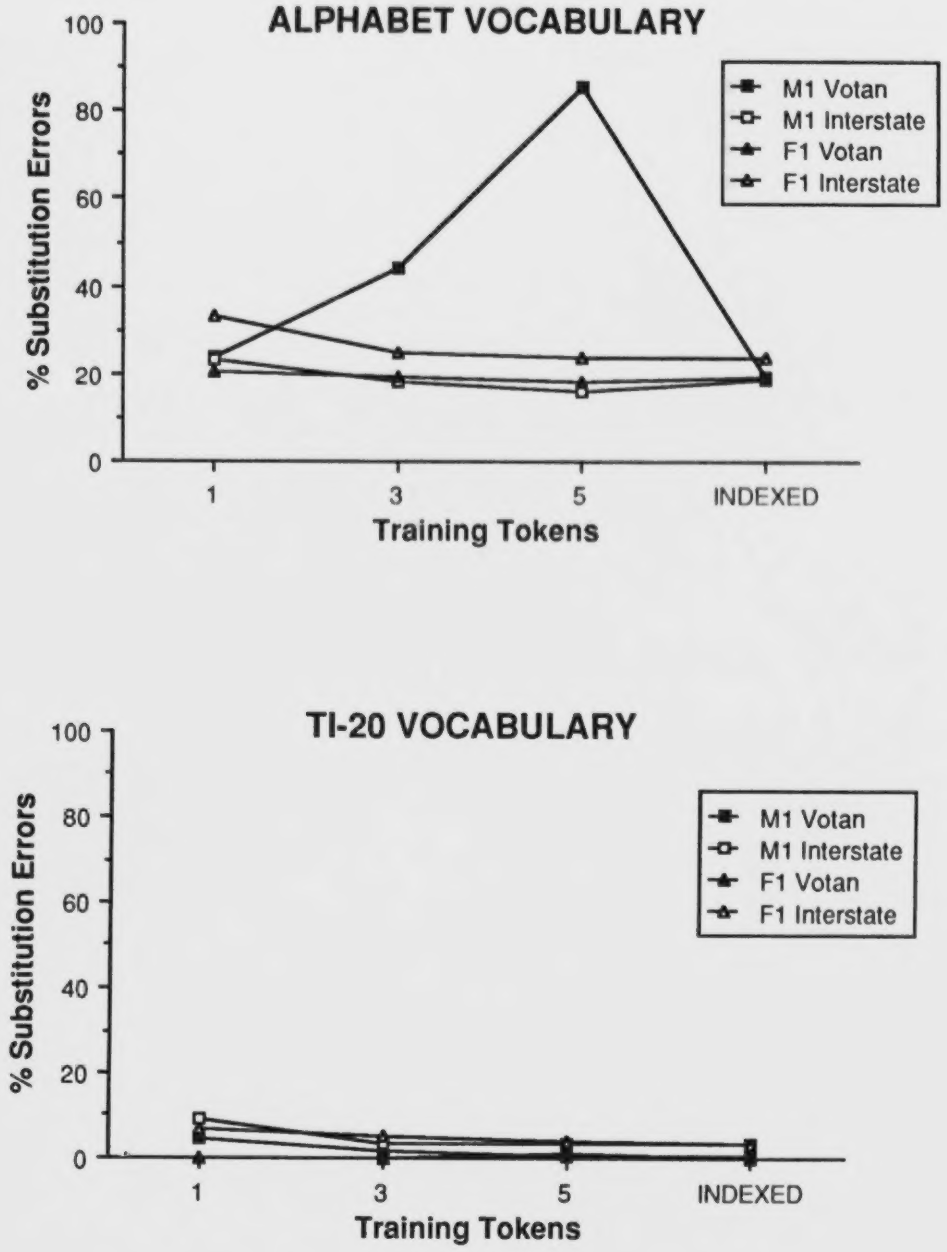


Figure 4. Recognition performance for the Interstate and Votan recognition systems tested on the speech of two specific talkers -- M1 and F1 -- for the Alphabet (top panel) and the TI-20 vocabularies (bottom panel). Note the substantial increase in error rate with increased training for the Votan system tested on M1 producing the Alphabet. The condition labeled *INDEXED* refers to maintaining separate vocabulary items for each training token.

variability to the recognition process.

Conclusions

Taken together, the results of our research on the performance of recognition devices demonstrate the importance of controlled laboratory testing. By carrying out tests with different amounts of training and different vocabularies, it is possible to draw some general conclusions about recognition performance of the six recognition systems we studied. First, it is clear that the difference in performance among these recognition systems is much smaller than would be expected based on the differences in technology and price. Second, the performance differences are largest when the recognition systems are inadequately trained. Thus, if a recognition device is trained appropriately, factors other than performance may be more important in selecting a recognizer for a specific application. In addition, it is important to note that appropriate training becomes much more important as the difficulty of the vocabulary increases. Finally, analyses of the performance data for one talker in the database suggest that the distinction between sheep and goats -- that is between successful and unsuccessful users of recognition devices -- may, in some cases, be the result of an interaction between talker, vocabulary, and a specific training algorithm. Eliciting new tokens, changing the vocabulary, or modifying software may reduce or eliminate the performance problem observed with these types of talkers.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 12 (1986) Indiana University]

Human Factors Issues for the Next Generation of Speech Recognition Systems*

Howard C. Nusbaum and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*Preparation of this paper was supported in part by NIH Research Grant NS-12179 and in part by Air Force Contract No. AF-F-33615-83-K-0501 through the Armstrong Aerospace Medical Research Laboratory (AFSC), Wright-Patterson AFB, OH to Indiana University in Bloomington. A version of this paper appears in the Proceedings of Speech Tech-86.

398

Abstract

There are a number of fundamental human factors issues that are concerned with the design and use of speech recognition systems. Some human factors problems are determined specifically by characteristics of the user, such as individual differences in the success of recognition performance. Other issues are more general, and relate more directly to the design of the user/system interface, such as the type of feedback presented to the user. While many of these issues will be the same for small and large vocabulary recognition systems, their relative importance and the way these issues are resolved will be very different in these two types of systems. This paper will outline some of these basic differences and will discuss some new human factors issues that may arise with the next generation of speech recognition systems.

Human Factors Issues for the Next Generation of Speech Recognition Systems

Introduction

Speech is the normal mode of communication between people. We can issue commands, requests, and assertions by directly speaking to another person, and although communication by speech is not a perfect process, we understand each other with a very high success rate. Furthermore, when an utterance is not understood completely, there are a number of standard linguistic conventions for error correction and recovery, such as a directed request for clarification of some part of an utterance. However, by comparison, speaking to machines is a much less satisfactory communication process. In general, talkers must pause between words, carefully choosing each word from a restricted set of alternatives and they must speak clearly and regularly. Since the interaction between human and machine is seldom an intelligent dialogue, feedback about errors and error correction strategies may be very frustrating to the user.

Clearly there is good reason to believe that speaking to a machine using current speech technology is not as simple as talking with a person. The act of choosing words from a relatively large but restricted vocabulary (greater than 50 but less than 20,000 words) and of speaking in a precise fashion may require a great deal of effort and attention. There is little doubt that the use of speech recognition systems in a variety of applications produces human factors problems that may not arise with other interface technology. Currently, some research has been directed at investigating the human factors problems that are involved in the use of commercially available, isolated utterance, small vocabulary speech recognition devices for database retrieval, command and control, and personnel training applications (see Simpson, McCauley, Roland, Ruth, and Williges, 1985). However, there has been very little research directed at the issues surrounding the use of large vocabulary recognition systems (e.g., Gould, Conti, and Hovanyecz, 1984) or the next generation of recognition systems still under development. Although vocabulary size has been cited as a significant limiting factor on the usability of speech recognition devices (NRC, 1984), it does not seem reasonable to assume that simply increasing vocabulary size will greatly enhance the performance or effectiveness of the technology. Rather, it is important to consider in detail the nature of the limitations on the technology and tailor the user interface to surmount those limitations wherever and whenever possible. Moreover, large vocabulary recognition systems will make possible new applications for speech technology such as dictation, and these applications will raise new human factors problems.

Our purpose in the present paper is not to attempt to describe the solutions to these problems, but instead to point out the need for new, systematic research that investigates the use of speech recognition systems both by making use of existing technology and by simulating more advanced systems (e.g., Gould et al., 1984). At the present time, there is simply too little information available about the human factors issues that must be addressed in developing and using the next generation of speech recognition systems. However, it is clear there are three aspects to the use of speech recognition systems that must be investigated: (1) the design and functions of the speech recognition system, (2) the limitations and capabilities of the human operator, and (3) the environment within which recognition takes place. Each of these three areas must be understood because there is little doubt that they interact to determine the overall effectiveness and performance of a speech recognition system for a particular task.

The Recognition System

For our purposes, a recognition system can be defined to consist of the hardware and software that instantiate the algorithms for training and recognition along with any software for user interaction with the recognizer's functions. There are actually two user interfaces that must be considered -- the recognizer interface and the application interface. At this point, our concern is with the information and control provided by the recognizer interface.

According to Norman (1983), a user interface not only serves to provide communication and control for a system, but it also conveys a system image. Since the user interface is the main point of contact between an operator and a system, it is through this interface that the user forms a mental model of the operation of the system. This mental model of a system allows the user to interact with, interpret, understand, and predict the behavior of the system. In a well-designed interface, the system image should directly reflect the operating principles of the system and should direct the user to interact with the system in clearly constrained ways. This issue of the development of a system image for a speech recognition device has not been explored systematically, perhaps because of the relatively simple operating principles for small vocabulary recognition systems. However, as the complexity of recognition systems increases, it will become very important to design interfaces that modulate the user's expectations about how to speak, about recognition performance, and about how to control the recognizer's functions.

Recently, Zoltan-Ford (1984) has demonstrated how the system image of speech technology can control the speech behavior of users. Subjects spoke freely with unconstrained syntax and vocabulary to a computer simulation of a recognition system that responded using a limited vocabulary and syntax. The subjects learned the syntactic and lexical constraints provided by the simulation, and adopted them in their own speech to the computer. Thus, it is apparent that the user does indeed form a mental model about speech technology through interactions with the user interface. Furthermore, by projecting the appropriate system image, it is possible to modify the speech produced by a user to conform to the requirements of a recognition system, thereby improving performance.

In addition, Norman (1983) has suggested that it is important to separate the user interface from the rest of a system. The reason for this is that a separate interface module makes it much easier to modify and improve the interface independent of the main functions of the system. In the case of a speech recognition device, it is important to provide the applications developer with the tools to build the recognition interface into the application. Since it is unlikely that the applications developer will be familiar with the requirements of a speech interface, it is also important to provide guidance in terms of how speech can be integrated into an application, along with examples of error correction procedures and feedback to the user. Although there has been some research about the type of feedback needed to improve the effectiveness of recognition technology (e.g., Schurick, Williges, and Maynard, 1985), it is unlikely that the results of this research will be appropriate for the next generation of recognition systems with large vocabularies. For example, it is unclear whether word-by-word visual feedback in a dictation task will slow down the dictation process or speed it up. Moreover, the type of error correction strategies that are effective for restricted vocabulary, data entry tasks will be less effective for large vocabulary dictation tasks.

Recognition performance is another important human factors issue that must be investigated systematically. Although it is clear that users will probably prefer a system with high recognition accuracy, it is not clear how users would trade off recognition accuracy for other features of a system. For example, would users accept lower accuracy with a very fast system response time? Or, will users prefer to dictate connected speech even if accuracy is lower than in a recognition system that requires isolated word input? Moreover, these issues can only be addressed in the context of a particular application.

In addition, it is clear that effective training of a recognition system is critical to the performance of that system. This issue becomes especially important for large vocabulary recognizers. There may be 5000 or more words in an isolated utterance recognition system, or hundreds of thousands of possible sentences in a connected speech recognition system, and it is not practical to require training on all the utterances that a device must recognize. As a consequence, it will be important to combine talker tuning of a recognition system (in which a talker provides a few key utterances prior to using the system) and talker adapatation (in which the representations of speech are changed on the fly following each recognition pass), in order to achieve acceptable performance. It is important to develop "tuning vocabularies" that provide critical information about talker differences and protocols for collecting the speech that insure that the talker does not speak in a different manner from the way the system will be used. Unfortunately, there has been little or no systematic research on this problem or on other human factors problems that stem from the size of the vocabulary in the next generation of recognition systems.

However, there has been one well-documented investigation of some of the issues surrounding vocabulary size that may arise with large vocabulary recognition systems. Gould et al. (1983) reported results of a dictation task based on a simulation of a large vocabulary recognition system. The results of this study may have important implications for several human factors issues. For example, they found that, with connected speech input, the speed of composing and proofing letters was independent of vocabulary size for users who were experienced in dictating letters, while there was a large effect of vocabulary size for isolated word input. For the experienced subjects, dictation performance became slower for isolated word input as the vocabulary size decreased. In general, for vocabularies of 5000 words or 1000 words, Gould et al. found that connected speech produced better performance of the users in writing letters. However, it is reasonable to suspect that a different pattern of results might be obtained for other applications. In any case, it is clear from this study that the type of speech input and vocabulary size are two issues that must be investigated more thoroughly.

Another issue of some importance for dictation tasks is the presence and weighting of a language model in a recognition system. On one hand, a language model that is built to constrain word recognition can improve performance in a listening typewriter as long as the speech input conforms to the expectations of the model. On the other hand, a highly weighted language model that is built on a particular database of text could actually introduce errors in recognition if the user's syntax and vocabulary was inconsistent with the model. It is important to investigate whether users can learn to accomodate tacit feedback from a language model or whether language models must be made to adapt to the talker over time.

Finally, it is important to consider how the user will change modes of operation. For example, in their simulation of a dictation application, Gould et al. (1984) provided several modes of operation to the user including a dictation mode (the default), a spell mode, a number mode, and several commands for formating and punctuating text. They provided clear keywords to enter and exit these different states. This is a very important human factors issue, since many errors in text editing occur because users may become confused about the current mode of a system. Norman (1984) has made four general suggestions about the design of a user interface for any complex information processing system. First, feedback to the user should provide clear information about the current state of the system. Second, different types of functions should be invoked using distinctly different commands. Third, the user should be able to undo almost all actions. Finally, it is important for commands to be as consistent as possible across the different modes of a system. For example, the delete function in a listening typewriter should be activated in the same way in dictation, spell, and number modes.

In general, few if any of these issues have been investigated systematically for the operation of large vocabulary speech recognition systems. However, it is important to realize that, although we have focused on the recognition system and its interface with the user, none of these issues can be completely understood without considering the limitations and capabilities of the human operator.

The Human Operator

The effectiveness of any recognition system will depend on the human operator who is providing the speech input and must make use of this technology. As a consequence, it is important to remember that the human information processing system is limited in its ability to perceive, encode, store, and respond to all the sensory stimulation that is presented at any point in time. Furthermore, the user is subject to stress and may become fatigued while using a recognition system. These limitations may, under certain conditions, constrain the performance that may be achieved with a recognition system. Thus, the overall effectiveness of recognition technology depends on a thorough understanding of the limitations of the human operator. However, at the same time, it is also important to remember that the human operator is much more flexible and accomodating than a speech recognition system. Humans are able to quickly adopt new strategies for interacting with technology if they are given appropriate and informative feedback. The human information processing system is adaptable and can respond to the requirements and limitations of technology when the technology is inflexible. Thus, even though there are limits to the human that form the boundaries on performance, there are also important capabilities that the human can use to cope with and overcome limitations on technology.

In order to fully understand and predict the performance of a speech recognition system in a particular application, it is important to understand the stress and cognitive demands placed on the human operator. At the present time, there is insufficient data on the effects of stress and effort on speech production. However, it is a general principle of cognitive and perceptual processes that as the demands on the human observer increase, performance decreases across a wide range of tasks. Thus, there is reason to believe that speech production, which is a skilled motor task, should be affected by these factors. There is clearly a need for understanding the effects of effort and attention on the acoustic-phonetic structure of speech (NRC, 1984).

Furthermore, no data have been reported on the effort required to dictate passages of text with isolated word input. Although Gould et al. (1984) reported data on the performance of talkers in a dictation task using discrete utterances as input, they did not report the effects of this requirement on the cognitive capacity of the talker compared to continuous speech input. It is entirely possible that the requirement of dictating discrete words may require a great deal of attention and effort so that performance may degrade over time at a faster rate than it would for continuous speech input. Similarly, the restriction of dictating using a limited vocabulary of 1000 to 5000 words or a constrained language model may require more effort than dictating from an unrestricted vocabulary or using unconstrained syntax.

However, it is possible that if such attentional effects are found, they may be overcome through appropriate training and experience with a speech recognition system. In our laboratory, we have found that human listeners can be trained to improve their ability to recognize synthetic speech generated by rule (Schwab, Nusbaum, and Pisoni, 1985). This demonstrates that the human listener is flexible enough to adopt new perceptual strategies to improve recognition performance with synthetic speech. It seems quite reasonable to predict that attentional demands may be reduced for interacting with speech recognition systems by giving the user appropriate training and experience. Furthermore, this training may be used to modify a talker's productions to improve recognition performance. The human operator is quite able to learn to use technology more effectively when given appropriate feedback and training. Thus, even if a particular operator does not use a speech recognition device effectively when first introduced to the technology, it may be possible to improve performance through a systematic program of training.

In general, talkers differ in their ability to use speech recognition devices. This is the basis for the distinction between "sheep" (successful users) and "goats" (unsuccessful users) that has been made for speech recognition systems (Doddington and Schalk, 1981). Of the possible accounts of this distinction among users, there are three explanations that seem most likely. First, sheep may be highly motivated to use the technology while goats are not. It is not clear precisely how motivation does affect performance, but it is easy to generate several possibilities such as mumbling, not responding to feedback, or not paying attention to the state of the system. Second, sheep and goats may differ in the acoustic-phonetic structure of their speech such that the utterances produced by goats are simply less distinctive and discriminable from each other compared to sheep. Finally, goats may be less consistent in their productions so that training utterances are very different from each other and from utterances provided as input when the recognition device is used in an application.

Recently, we have obtained some data in our laboratory that bears directly on these hypotheses. We have been testing the performance of several small vocabulary, speech recognition systems using the digitized speech database collected by Doddington and Schalk (1981). We have found some systematic differences between talkers for performance on different recognition devices. Since the same recorded tokens are used for testing the different systems, motivation of the speakers is probably less a factor than the acoustic structure of the speech. However, a much larger difference between talkers was observed that depended on the recognition systems themselves. In the most striking example of this finding, as the number of training tokens increased, the pattern of performance for one talker was very different across two recognition systems. For one system, as the number of training tokens was increased, recognition accuracy increased as would normally be expected. But for a different device, as the number of training tokens increased, recognition accuracy actually *decreased*. Furthermore, this pattern was only found for this talker for an alphabet vocabulary and not for a digit vocabulary. This finding suggests that the definition of a goat may depend on the training algorithm, the recognition algorithm, *and* the specific vocabulary used for recognition. This indicates that the issue of goats and sheep may be less a function of the talker and more a function of the recognition technology itself. However, it is apparent that differences in the speech produced by different talkers must be understood so that it is possible to modify speech behavior through training and experience to improve recognition performance.

The Recognition Environment

Finally, speech recognition takes place within the context of a physical environment and an application environment. The physical environment includes the ambient noise background, as well as other nonacoustic characteristics such as temperature, vibration, and acceleration. These physical characteristics can affect recognition performance by either changing the way the talker produces speech or by affecting the input or operation of the recognition system. In particular, the effects of high levels of background noise on recognition performance are well known (NRC, 1984). However, recently we have demonstrated that the acoustic-phonetic properties of speech are modified for speech produced in noise, compared to speech produced in quiet (Pisoni, Bernacki, Nusbaum, and Yuchtman, 1985). In essence, there is a systematic change in the tilt of the power spectrum when there is noise presented to the ears of the talker, along with systematic changes in the vowel space. These changes can have large effects in reducing recognition performance. As a consequence, recognition systems will have to adapt to changes in a talker's speech that might occur as a consequence of a wide range of environmental factors.

Beyond the human factors problems engendered by conditions of the physical environment, it is also important to understand the effects of the application environment on the effectiveness of using speech recognition systems. The application environment refers to the tasks and application context in which speech recognition is being used. It has been stated many times that the appropriate applications for speech input to replace manual input are tasks in which an operator's eyes and hands are already busy. Speech input is often proposed to provide a more efficient control system, or to provide another set of functions that is not available in the existing system. As a general rule, this is probably not a bad guide for small vocabulary recognition systems (e.g., Simpson et al., 1985).

However, this may not be the most important criterion for using large vocabulary, speech recognition systems. For these new, more powerful systems, it will be important to consider tasks in which speech is currently used with a human who is listening and responding in some way. In these cases, a speech recognition system would directly replace the functions carried out by a human listener. The prototypical example of this is dictation in which text is spoken to a secretary for transcription. However, there are other examples as well, such as database retrieval in which queries may be very complex and might be made over a phone to a database manager. To be effective in these applications, speech recognition systems must not interfere with carrying out the task. There should be little difference in the demands placed on the user whether a human or a machine is listening.

Also, speech recognition should facilitate the task in some way that is readily apparent to the user. In considering the use of small vocabulary recognition systems for voice data entry tasks, a significant advantage is the immediate entry of data into a computer system for higher level managers to access and analyze. The benefits to the user are not direct in this case and the operator is instructed to use voice data entry as part of the job. However, in the case of an application like dictation, the user must be motivated to change from a human listener to a machine.

Moreover, there is a need for research on the relationship between the task requirements of different applications and the functions that might be provided by a large vocabulary speech recognition system. There are currently no strong criteria for choosing a particular application as a target for speech recognition. Instead, recognition is integrated into applications based on a combination of intuition and trial and error. Furthermore, there are no guidelines to specify how to integrate recognition functions into an application. With large vocabulary recognition systems, there may be a temptation to place all the functions of the application under the control of the recognizer. However, this may not be the best approach for all applications. For example, in a dictation task, speech may provide the best means of entering the original text into the system, but it may be more efficient to edit the text using a keyboard and a mouse.

Summary and Conclusions

Although there has been some research in recent years directed at investigating human factors problems in using speech recognition systems, this research has focused primarily on small vocabulary systems and the applications that are appropriate for this more limited technology. However, as the capabilities of recognition systems increase, so does their complexity. The next generation of speech recognition systems will be used in very different applications and will place very different demands on the human operator. In order to understand better how to use these systems effectively, it is important to investigate more thoroughly a number of human factors issues.

There are three constraints that must be understood in order to optimize the use of speech recognition systems -- the characteristics of the recognition system, the limitations and abilities of the human operator, and the requirements and demands of environment. These three general constraints interact to determine the overall performance of a particular recognition system and there can be little doubt that as the complexity of recognition systems increases, so will the importance of understanding the human factors problems inherent in each constraint.

References

Doddington, G. R., and Schalk, T. B. (1981). Speech recognition: Turning theory to practice. *IEEE Spectrum*, *18*, 26-32.

Gould, J. D., Conti, J., and Hovanyecz, T. (1984). Composing letters with a listening typewriter. *Communications of the ACM*, *26*, 295-308.

National Research Council (1984). *Automated speech recognition in severe environments*, Committee on Computerized Speech Recognition Technologies, Washington, D.C.

Norman, D. A. (1983). In *Proceedings of the CHI 1983 Conference on Human Factors in Computer Systems*, Boston, December.

Norman, D. A. (1984). Design rules based on analyses of human error. *Communications of the ACM*, *26*, 254-258.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., Yuchtman, Y. (1985). Some acoustic-phonetic correlates of speech produced in noise. In *Proceedings of ICASSP 85*. New York: IEEE Press.

Schurick, J. M., Williges, B. H., and Maynard, J. F. (1985). User feedback requirements with automatic speech recognition *Ergonomics*, *28*, 1543-1555.

Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*, 395-408.

Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., and Williges, B. H. (1985). System design for speech recognition and generation. *Human Factors*, *27*, 115-141.

Zoltan-Ford, E. (1984). Reducing variability in natural language interactions with computers. In *Proceedings of the Human Factors Society 28th Annual Meeting* (Vol. 2), Santa Monica, CA: Human Factors Society.

Using Speech as an Index of Alcohol-Intoxication*

Christopher S. Martin and Moshe Yuchtman

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This paper is a summary of a poster that was presented at the 111th meeting of the Acoustical Society of America in Cleveland, Ohio, May 12 - 15, 1986. This work was supported by a contract from General Motors Research Laboratories to Indiana University in Bloomington. We thank Susan Hathaway for her help in the early phases of this project. We also thank Dr. Robert Levenson for his help with this project and the staff of the Psychophysiology Laboratory, Sandi Houshmand and Jeni Hayes, for their assistance in preparing talkers for the recording sessions. We would especially like to acknowledge the help and assistance of the Indiana State Police, Bloomington Post. Without their cooperation, this study would not have been possible.

Abstract

In a recent study in our laboratory, we found that listeners could reliably discriminate between sentences produced by talkers speaking under sober and alcohol-intoxicated conditions, and that experienced listeners performed significantly better than naive listeners (Pisoni, Hathaway, and Yuchtman, 1985). The present experiment was designed to study how well listeners can distinguish between speech produced in a sober and an alcohol-intoxicated condition, using a single-interval absolute identification task. Subjects were required to make a judgement for each sentence in isolation, rather than a comparative judgement on a matched pair of sentences. Two groups of listeners, college students and Indiana State Troopers, were tested in order to evaluate the effects of past experience in detecting changes in speech due to alcohol-intoxication. We were interested in determining whether speech could be used as an index of sensory-motor impairment due to alcohol-intoxication. Both groups of listeners identified test sentences significantly above chance for all eight talkers used in the experiment. State Troopers performed significantly better than college students for six of the eight talkers. The results demonstrate that systematic changes in sensory-motor control are encoded in the speech waveform, and that listeners can reliably identify these properties in a single-interval absolute identification task. There also appear to be reliable differences between groups of listeners as a function of their experience in detecting these changes in the speech waveform.

Using Speech as an Index of Alcohol-Intoxication

Although laypersons and law-enforcement officers routinely use perceived speech quality as an index of alcohol-intoxication, little systematic research has been done on the accuracy and limitations of this ability. The few studies that have examined differences in perception between speech produced in a sober and in an alcohol-intoxicated condition are consistent with the finding that alcohol acts as a central nervous system depressant that effects fine sensory-motor processes. Speech produced in an alcohol-intoxicated condition has been found to be more prone to errors, is lower in amplitude, and is more negatively judged in perceptual tests. These findings were discussed in Pisoni, Hathaway, and Yuchtman (1985).

Data for the present experiment were collected by Pisoni, Hathaway, and Yuchtman (1985). They made audio recordings of eight male talkers in a sober and an alcohol-intoxicated condition. Blood alcohol level (BAL) was determined by a Breathalyzer test. In a perceptual experiment using an A-B forced choice format, graduate students, versed in articulatory phonetics, and introductory psychology students both performed better than chance in identifying the condition under which the sentences of four male talkers were produced. Subjects listened to the same speaker produce the same sentence in both conditions, and selected the sentence produced in the intoxicated condition. Overall percentage of accuracy was 82.4% for the graduate students, and 73.8% for the introductory psychology students, suggesting that listeners who have studied speech can use perceptual cues in the speech waveform as an index of alcohol-intoxication more accurately than listeners who have not had this experience. The absolute frequency of obvious misarticulation errors in these sentences was very low, and could not account for most of the listeners' discriminations.

The present experiment was designed to study how well listeners could distinguish between speech produced in a sober and an alcohol-intoxicated condition with a single-interval absolute identification task. Subjects were required to make an absolute judgement for each sentence in isolation, rather than a comparative judgement on matched pairs of sentences used in the previous perceptual experiment. Two groups of listeners, college students and Indiana State Troopers, were tested in order to evaluate the effects of past experience in detecting changes in speech due to alcohol-intoxication on this ability. We were interested in determining whether speech could be used as an index of sensory-motor impairment due to alcohol-intoxication.

Method

Subjects Two groups of subjects were used. One group consisted of thirty introductory psychology students who received credit to fulfill a course requirement. The second group consisted of fourteen Indiana State Troopers who volunteered their time to participate in this study. All subjects were native English speakers with no history of a speech or hearing disorder.

Stimuli Speech samples from eight male talkers speaking in a sober and an alcohol-intoxicated condition were recorded by Pisoni, Hathaway, and Yuchtman (1985). Two master files of sentences digitized at 10 KHz were compiled from this data bank. Each file contained eight talkers speaking the same 24 sentences. Each talker contributed 12 sentences produced in an intoxicated condition, and 12 different sentences produced in a sober condition. The two master files differed in that each sentence-speaker combination appeared only in the intoxicated condition for one file and the sober condition for the other. Different random orders of each master file were transferred from the computer onto audio tapes using a 12-bit digital-to-analog converter. A five second interval was inserted between the sentences. Half of the subjects in each group heard an audio tape generated from the first master file, and half heard an audiotape generated from the second master file.

Procedure Each listener heard a total of 192 sentences, eight talkers saying 24 sentences each. Subjects wore headphones and were presented with the sentences on audio tape. They recorded their decision after each sentence by circling the letter S for "sober" or I for "intoxicated" on a prepared response sheet and then rated their degree of confidence in their choice on a scale from 1, "just guessing", to 5, "very sure."

Results

Accuracy Mean accuracy across all of the sentences was 61.48% (SD=4.25) for the college students, and 64.66% (SD=3.08) for the Indiana State Troopers. This difference was significant (t=2.43, p<.02). Both groups performed significantly better than chance beyond the .001 level. Mean accuracy for the sentences produced in the two conditions was 60.5% for sober, and 64.5% for intoxicated. This difference was significant beyond the .01 level. These data are shown in Figure 1.

Insert Figure 1 about here

Mean accuracy for the different talkers ranged from 55% to 71.9%. Accuracy for all of the talkers was significantly better than chance beyond the .001 level. State Troopers performed significantly better than college students for 6 of the 8 talkers. These data are shown in Figure 2.

Insert Figure 2 about here

No differences in performance between the groups for sentences produced in the intoxicated condition were observed for any of the eight talkers. However, State Troopers performed significantly better than the college students in correctly identifying sentences produced in the sober condition for 6 of the 8 talkers (p<.001).

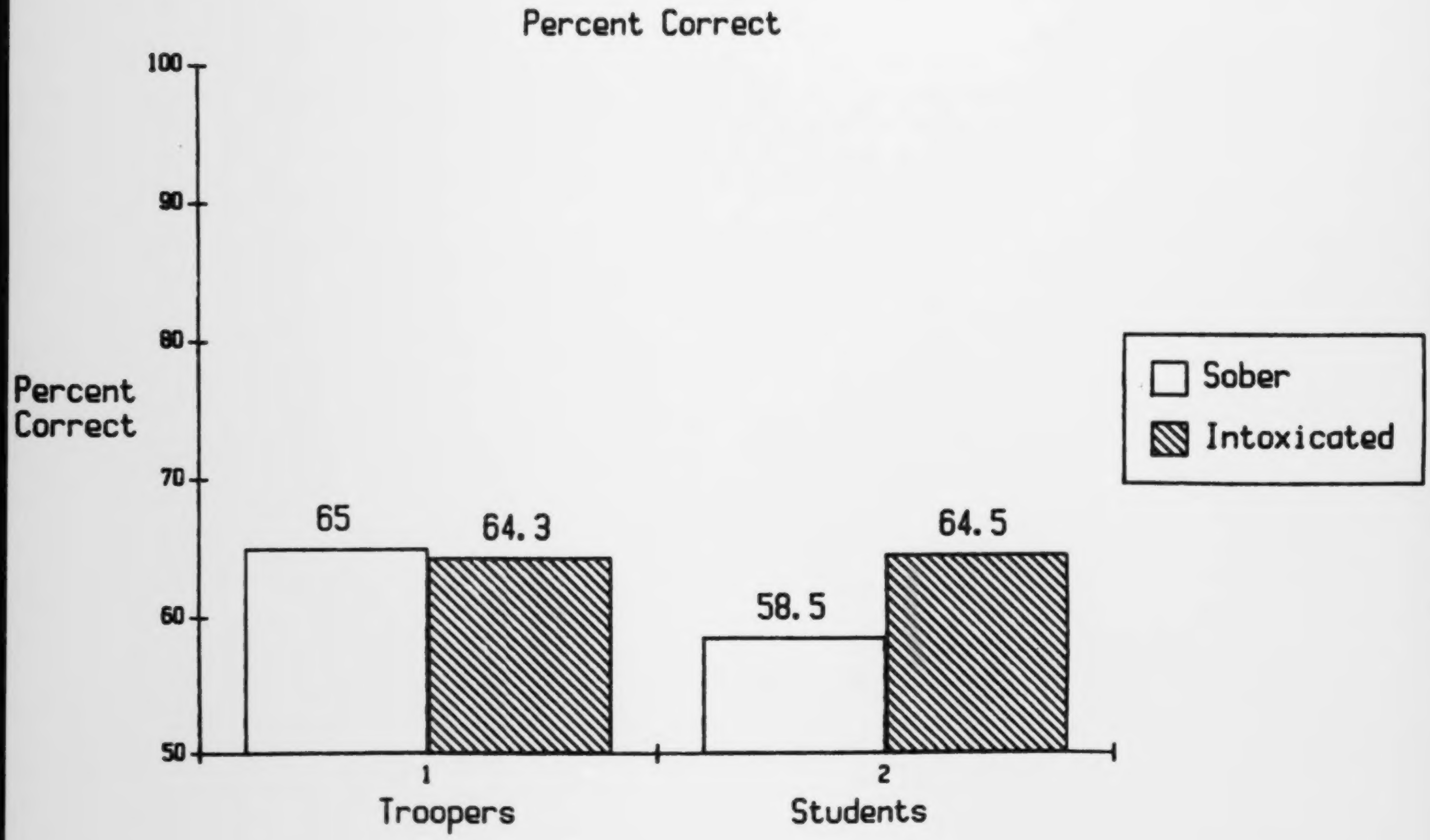


Figure 1. Mean percentage of correct responses by condition and listener group.

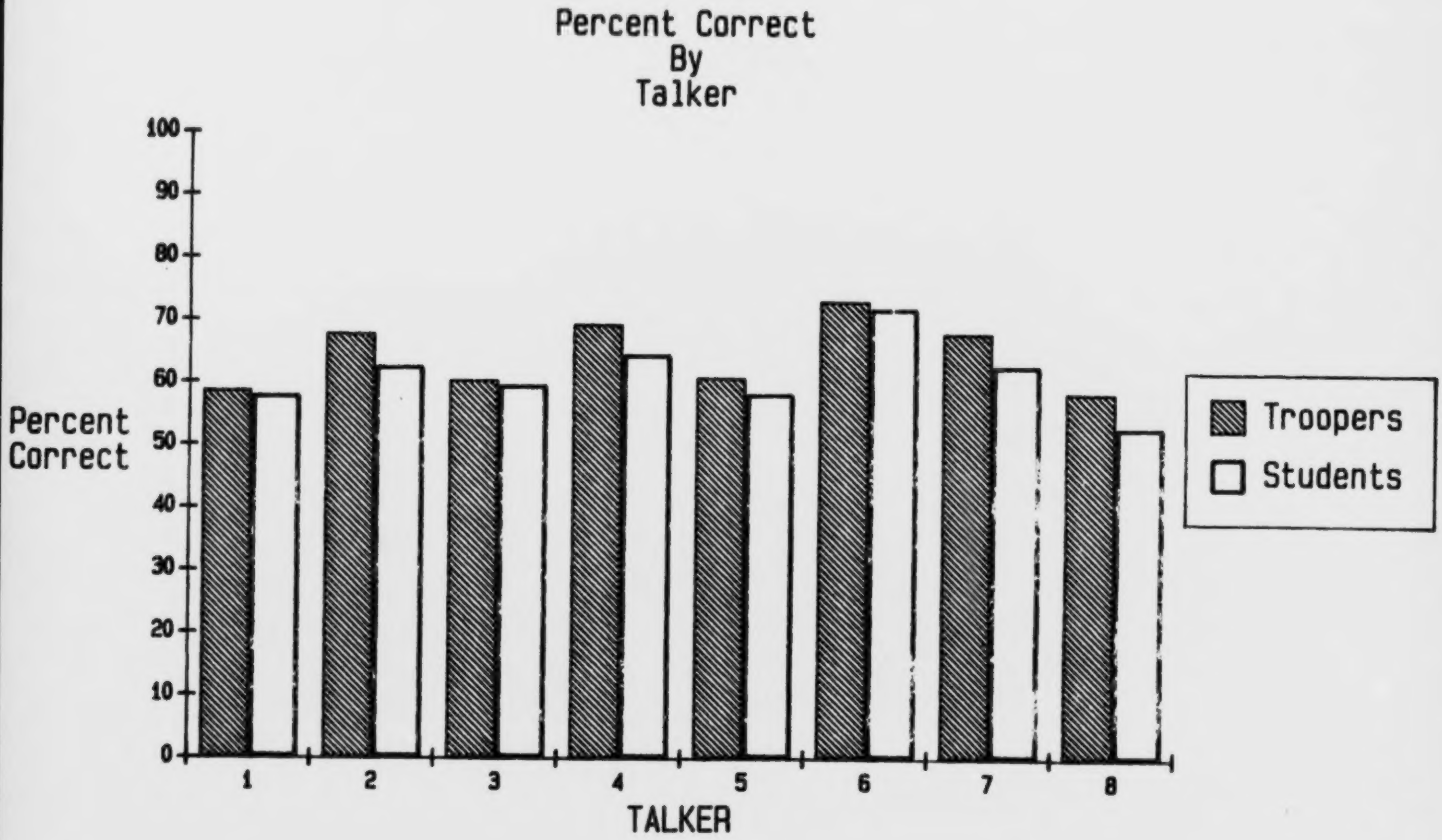


Figure 2. Mean percentage of correct responses by talker and listener group.

Response Bias Mean beta and d′ values for the two groups were calculated from the proportion of hits (correctly identifying a sentence produced in the intoxicated condition) and false alarms (identifying a sentence actually produced in the sober condition as having been produced in the intoxicated condition). Beta is a measure of response bias, and d′ is a measure of performance independent of response biases. Beta was slightly higher for State Troopers compared to students, but this difference was not significant. State Troopers obtained a significantly higher d′ compared to students ($p<.01$). These data are shown in Figure 3.

Insert Figure 3 about here

The proportion of hits and false alarms for both groups of listeners on individual talkers is plotted on an ROC graph in Figure 4. Performance for the individual talkers was very similar for both groups, although the State Troopers had a stricter criterion in judging whether a sentence was produced in an alcohol-intoxicated condition, for 7 of the 8 talkers.

Insert Figure 4 about here

The probability of hits and false alarms for individual talkers, ranked in ascending order of false alarm rate, is shown in Figure 5. There was considerable talker variability which led the speech of some talkers to be consistently labelled sober or consistently labelled intoxicated in both conditions.

Insert Figure 5 about here

Confidence Ratings The proportion of the total responses in each confidence rating category for both groups of listeners in shown in Figure 6. State Troopers used the extreme ratings of 1 and 5 less often, and a rating of 3 more often, compared to the students. Percent accuracy across the five confidence rating categories is shown in Figure 7. The more confidence a listener placed in a given response, the more likely it was that the response was correct. This finding suggests that listeners are able to reliably predict their ability to discriminate between the speech samples produced in the two conditions.

Insert Figures 6 and 7 about here

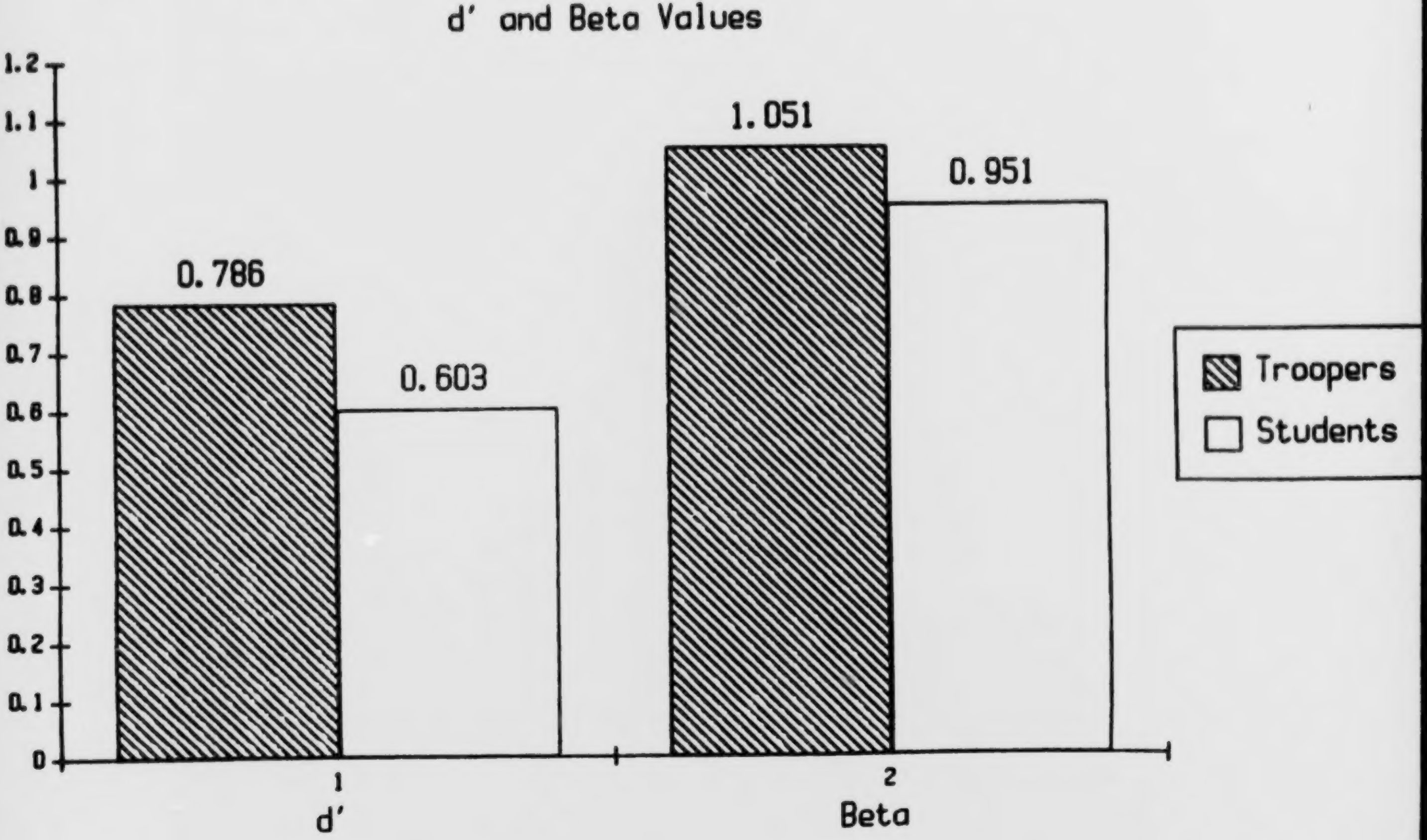


Figure 3. d' and beta values for the two groups of listeners.

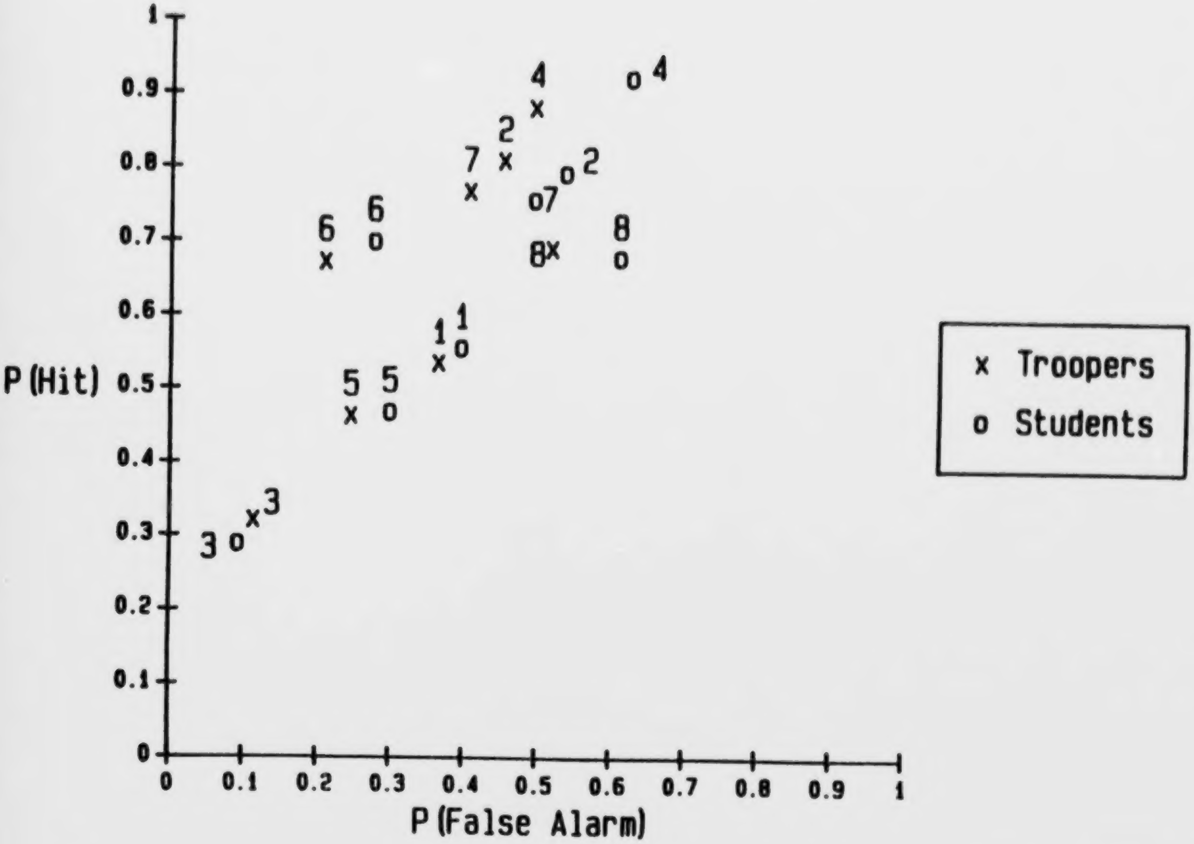


Figure 4. ROC graph for individual talkers by listener group.

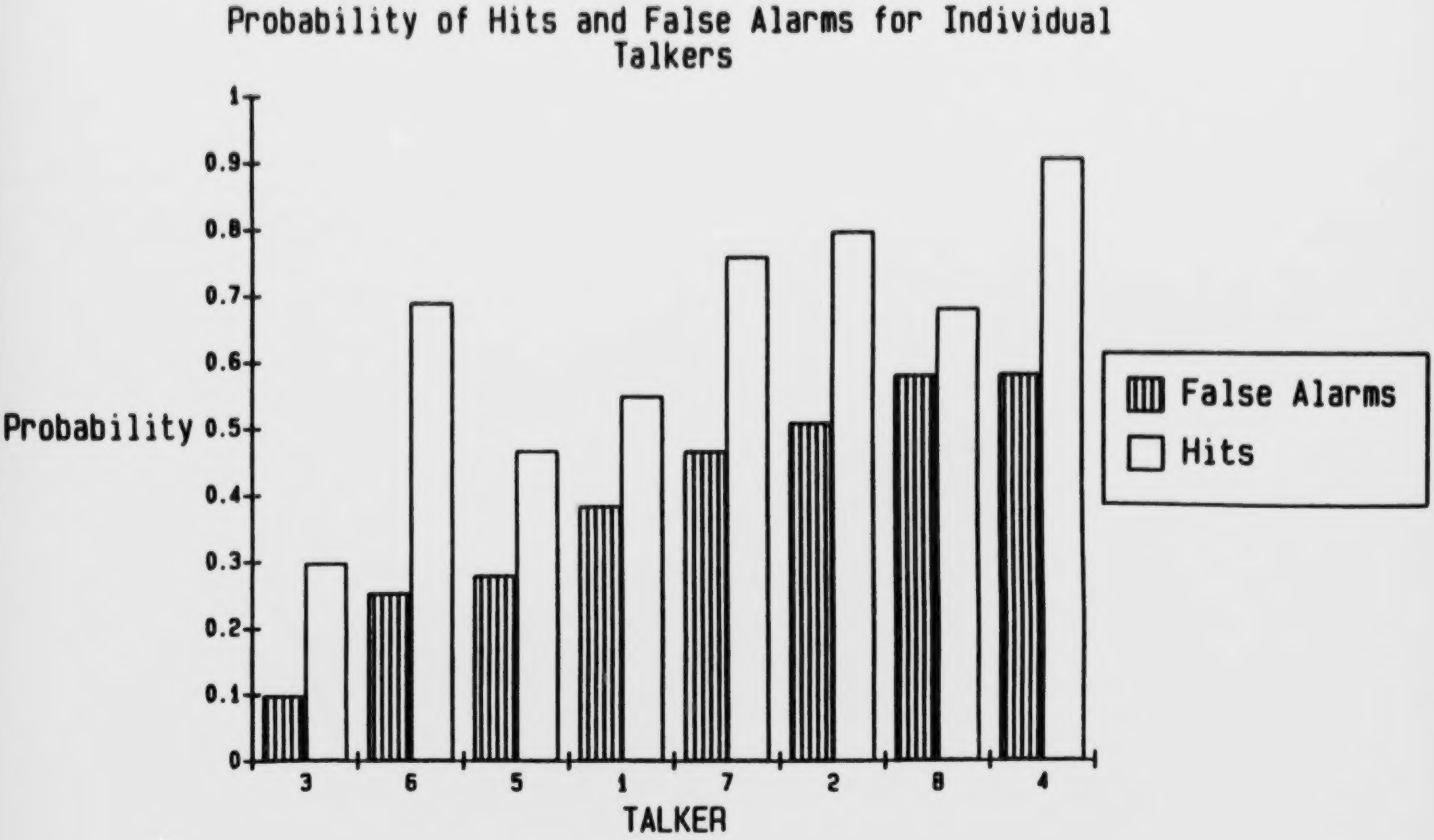


Figure 5. Probability of hits and false alarms by talker.

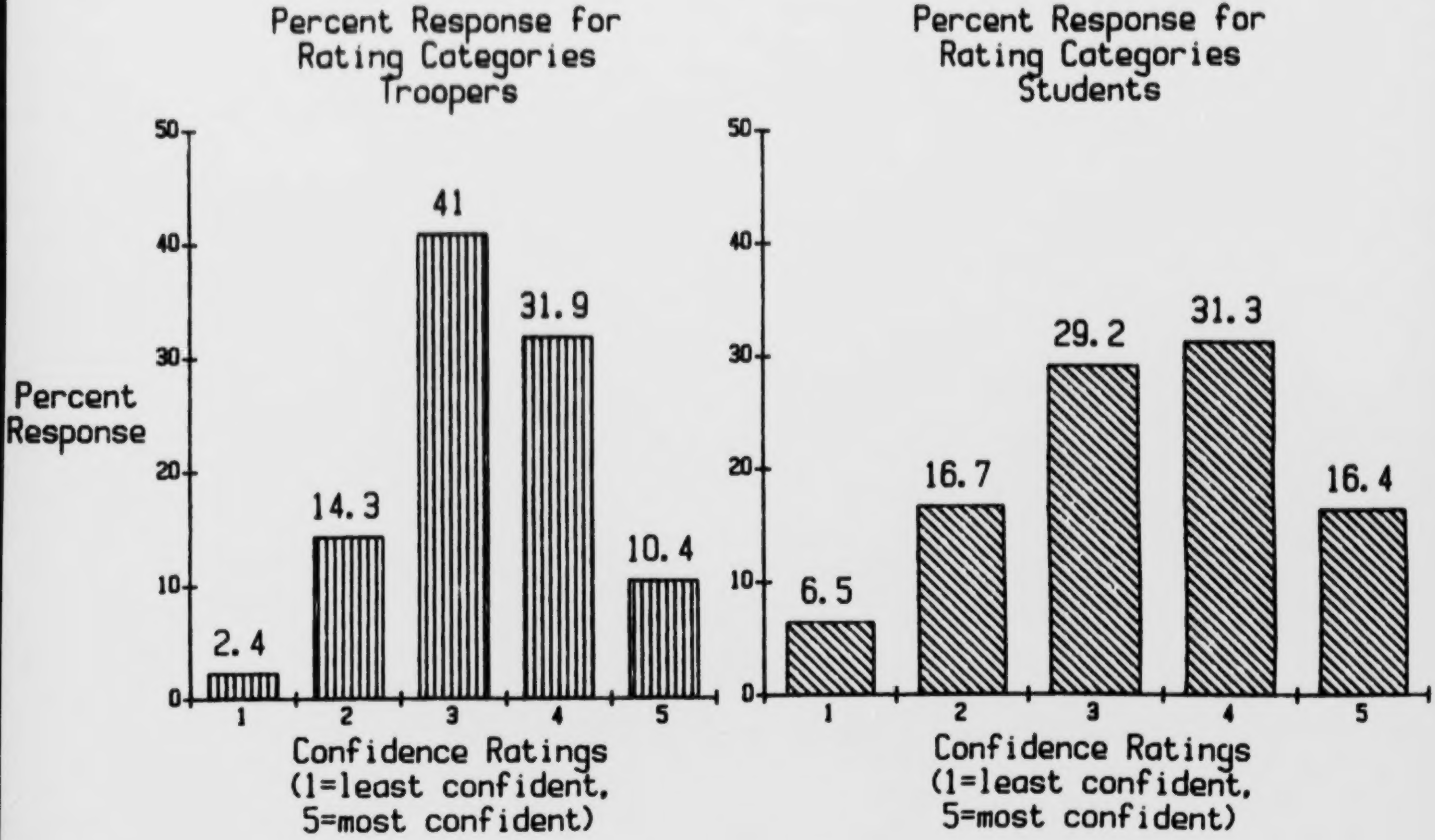
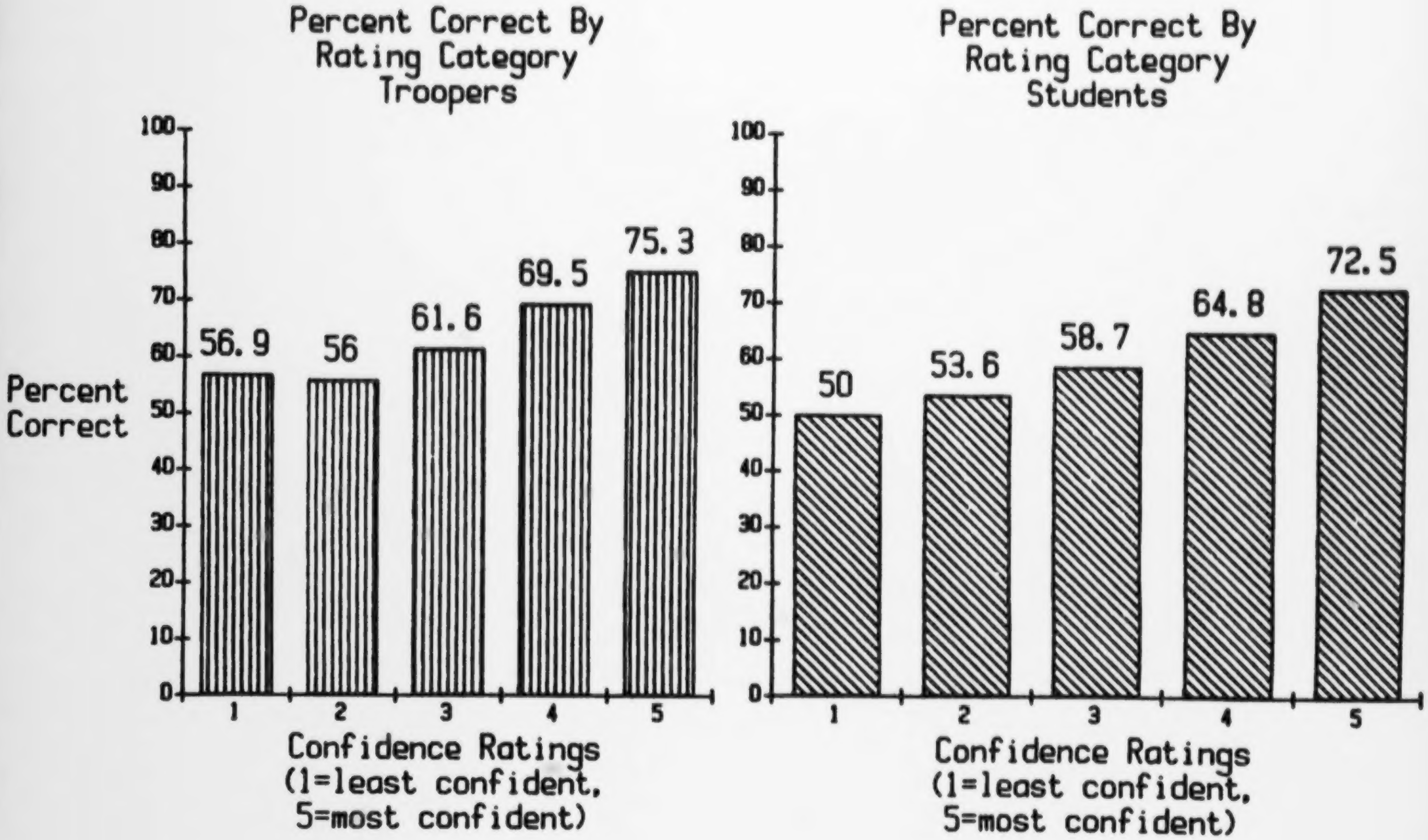


Figure 6. Proportion of total responses in each confidence rating category.


Percent Correct By Rating Category Troopers

Percent Correct By Rating Category Students

Figure 7. Percent correct for each confidence rating category.

Discussion

Our results support the hypothesis that perceived speech quality can be a reliable index of sensory-motor impairment due to alcohol-intoxication. Both groups of listeners performed significantly above chance level in identifying test sentences. However, no listener obtained higher than a 70% correct response rate, suggesting that the ability to use speech as an index of alcohol-intoxication may be limited when speech samples of short duration are used to make absolute judgements. The performance of listeners in discriminating the condition under which longer passages of fluent speech are produced needs to be studied. Percent accuracy for the different talkers ranged from 55% to 71.9%, suggesting that the degradation of speech quality varies widely across individuals with similar blood alcohol levels.

Our results also lend support to the hypothesis that experience in detecting changes in speech due to alcohol-intoxication enhances performance in the experimental task. State Troopers performed better than college students at this task. The difference between the two groups of listeners, while statistically significant, was quite small. It may be that the experimental task was not powerful enough to detect larger differences between the groups of listeners that actually exist, or that the two groups are more similar in their ability to use speech as an index of alcohol-intoxication than was hypothesized. Previous results in our laboratory suggest that experience in studying speech, compared to experience in detecting changes in speech due to alcohol-intoxication, may be a more important factor in using speech as an index of alcohol-intoxication. Obviously, further research is needed to investigate this issue.

State Troopers were significantly better at this task overall because they identified sober, and not alcohol-intoxicated, sentences better than the college students. Beta, a measure of response bias, was slightly higher for the State Troopers. Thus, the State Troopers used a stricter criterion in judging if a sentence was produced in an alcohol-intoxicated condition. College students were more likely to judge a sentence as being produced in an alcohol-intoxicated condition when it was not. The nonbiased measure of discriminability, d′, was significantly higher for the State Troopers compared to college students, suggesting a true difference in sensitivity. This conclusion was supported by the confidence ratings given by the listeners, which were highly correlated with response accuracy. The more confidence a listener placed in a given response, the more likely it was that the response was correct.

The present results demonstrate that systematic changes in sensory-motor control are encoded in the speech waveform and that human listeners can reliably identify these properties in a single-interval absolute identification task. There appeared to be reliable differences between groups of listeners as a function of their experience detecting these properties in speech. However, there were also substantial individual differences among the talkers who produced the speech samples. Some talkers were consistently labelled sober or consistently labelled intoxicated no matter what the true condition was.

Based on these findings and our earlier acoustical analyses, it may be possible to develop talker-dependent algorithms to identify sensory-motor impairment from speech samples, and to use these algorithms in safety interlock systems to prevent alcohol-intoxicated drivers from starting their automobiles.

References

Andrews, M. L., Cox, W. M., and Smith, R. G. (1977). Effects of alcohol on the speech of nonalcoholics. *Central States Speech Journal*, *28*, 140-143.

Lester, L., and Skousen, R. (1974). The phonology of drunkenness. In A. Bruck, R. A. Fox, and M. W. Lagaly (eds.), *Papers from the Parasession on Natural Phonology.* Chicago: Chicago Linguistic Society.

Pisoni, D. B., Hathaway, S. N., and Yuchtman, M. (1985). Effects of alcohol on the acoustic-phonetic properties of speech: Final report to GM Research Laboratories. *Research on Speech Perception Progress Report No. 11*. Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Sobell, L. C., and Sobell, M. B. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, *15*, 861-868.

Sobell, L. C., Sobell, M. B., and Coleman, R. F. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica*, *34*, 326-333.

Effects of Wholistic Versus Dimensional Training on Learning to Identify Spectrographic Displays of Speech*

Beth G. Greene

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

*This research was supported by NSF Grants BNS 80-23099 and BNS 83-05387 to Indiana University in Bloomington. I would like to thank David Pisoni and Thomas Carrell for their advice and assistance throughout this research project.

Abstract

Two groups of subjects learned to recognize speech spectrograms of 50 monosyllabic English words spoken by two male talkers. Ten blocks of five words each served as the stimulus words in this experiment. These word lists were prepared so that each block shared some phonetic attribute, e.g., initial /br/ cluster (brain) or the vowel-consonant pattern /Il/ (bill). The dimensional group viewed the spectrograms in a blocked condition; that is, they learned the spectrograms in blocks that shared a phonetic attribute. The wholistic group viewed five stimulus spectrograms selected randomly from the set of 50 words. Subjects learned to recognize the 100 spectrograms after 18 hourly sessions at 90% correct or better. When presented with the same word spoken by a new talker, subjects in the wholistic group correctly identified 74% of the spectrograms while subjects in the dimensional group correctly identified 58% of the spectrograms. The implications of these results for perceptual learning of visual displays of speech are discussed in terms of training procedures and the acquisition of detailed acoustic-phonetic knowledge about speech spectrograms.

Subjects were trained to recognize visual displays of speech as an English word using a study-test procedure. In this procedure, the subject viewed a spectrographic display of speech on a CRT screen and was told what the word was. The subject was instructed to "study" the display and to learn the relationship between the visual pattern and the word it represented. Subjects were also told that words that sound like each other look alike in spectrograms and words that sound different from each other look different from each other. After learning a predetermined number of stimulus items, the subject was given a "test" to evaluate how well the stimulus items were learned.

Subjects viewed a single spectrogram of each word displayed in the center of a CRT monitor screen. On those trials designated as "study" or "practice" trials, feedback was given verbally by the experimenter. Before feedback was provided, subjects were required to write down a response on prepared response sheets for every spectrogram displayed. On trials designated as "test" trials, no feedback was given. Subjects were required to respond on every trial even if they had to guess. All responses were recorded on prepared response sheets and saved for later analysis.

Stimulus Materials. In this experiment, the stimulus materials consisted of 50 monosyllabic English words. The words represented 10 phonetic dimensions, 5 words per dimension. The dimensions included shared initial or final phonemes (e.g., feed, fail, fan; bill, dill, hill); shared consonant clusters (e.g., bran, braid, broke; band, hound, land); and same initial and final consonants with the medial vowel varying (e.g., fizz, fours, fuzz; bid, bed, booed). Examples of the dimensions used in this study are listed in Table 1. Figures 1 and 2 show spectrograms of five words representing the /fVz/ and /-at/ dimensions.

Insert Table 1 and Figures 1 and 2 about here

Stimuli were presented to subjects in blocks of five words. One group of subjects, the dimensional group, received all five tokens of one dimension in a block. The first block of items for this group consisted of the items bill, dill, hill, mill, and frill. A second group of subjects, the wholistic group, received a random set of five words selected from the set of 50 words. For this group the items were bran, feed, find, hat, and frill. In addition, for both groups, we included tokens spoken by two different talkers. Thus, for each block of five words, ten spectrograms were presented to subjects, two for each word.

Each stimulus word was spoken in citation form and recorded on audio tape for later editing. Two males and one female produced tokens each of the 50 training words as well as tokens of 50 test words. All stimulus words were then processed digitally using a 12 bit analog-to-digital converter, edited into individual stimulus files, and stored for future use. Automated experimental programs were used to present blocks of spectrograms to subjects each day.

Table 1

Examples of Dimensions*

1. -ILL	BILL	CHILL
2. -EAR	GEAR	REAR
3. -AT	BAT	CAT
4. -ASH	MASH	STASH
5. F-	FACE	FLAT
6. BR-	BRAWL	BREED
7. -S	PASS	MICE
8. -ND	FIND	SEND
9. BvD	BID	BEAD
10. FvZ	FUZZ	FEES

---

*Dimensions are phonetic, not orthographic

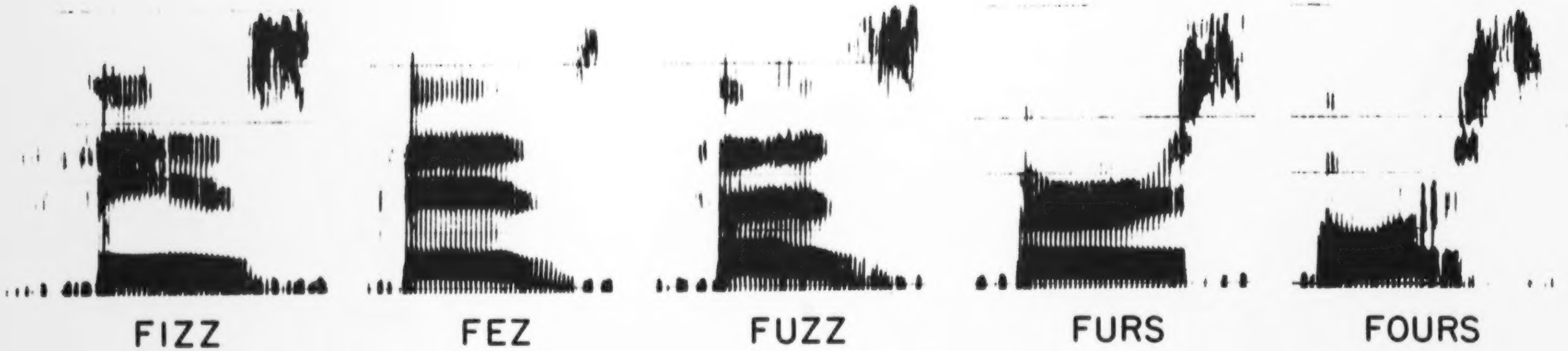


Figure 1. Speech spectrograms of five words for the /fVz/ dimension.

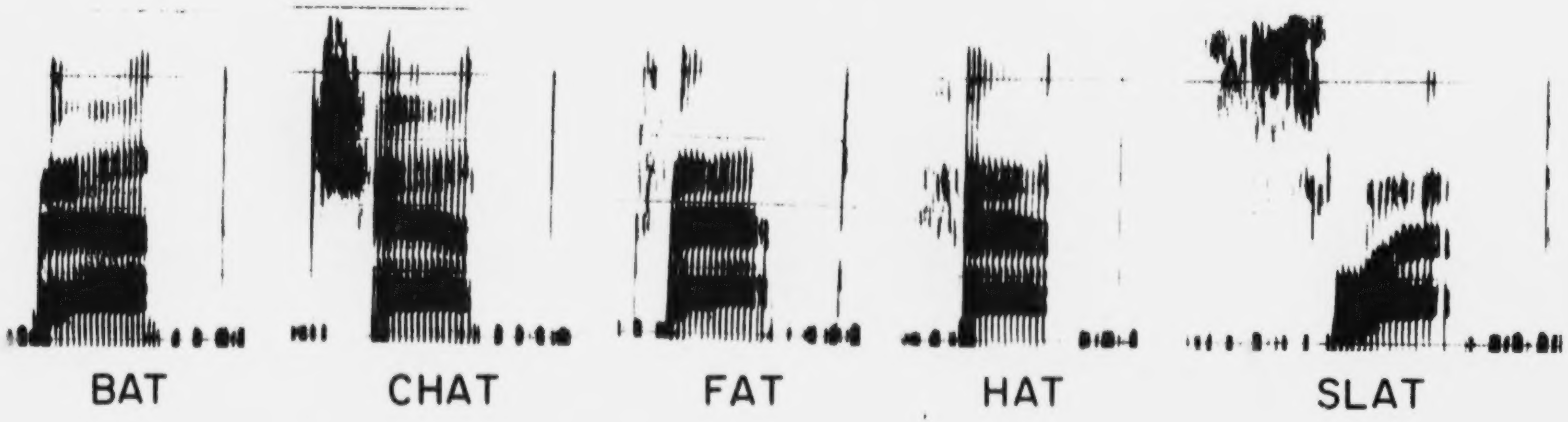


Figure 2. Speech spectrograms of five words for the /-at/ dimension.

Procedure. Each spectrogram was displayed in the center of the CRT screen for study. The display remained on the screen for five seconds after which the experimenter told the subjects what word the spectrogram represented. The spectrogram was then erased from the CRT screen. Subjects studied five words during each session designated as a study session. Since two different talkers produced tokens for this experiment, there were ten new tokens presented during each study session. Each token was presented five times during the session. Thus, each study session consisted of fifty trials.

Results

Subjects learned to recognize the 50 words in the training set at levels above 85% correct in 18 sessions. There was a total of just over nine hours of training and testing time during the course of the 18 day training period. The eighteen daily sessions were a maximum of one hour long but the actual study time, i.e., presentation of the spectrograms with feedback, averaged 10 to 12 minutes per session. The testing portion of the daily sessions took longer, ranging from 10 to 40 minutes. After each block of five new items was presented to subjects using the training procedure described above, subjects were tested for recognition on the new items plus all previously learned items. Therefore, the daily recognition test increased by 10 tokens each day. During these tests, no feedback was ever provided. Subjects had to identify each stimulus independently from every other on each trial since they were not told whether their response was correct. The daily test results are displayed in Figure 3.

Insert Figure 3 about here

As shown in Figure 3, subjects' performance never fell below 85% correct; in fact, performance was generally above 90% correct. Subjects in the wholistic group showed consistently higher levels of performance than subjects in the dimensional group throughout the entire experiment.

In contrast to our earlier study on reading spectrograms of phonetically balanced words, we did not provide extensive practice and testing prior to generalization testing as we did in our previous study (Greene, Pisoni, & Carrell, 1984). After the subjects had learned the 50 words in the training set, we went directly to generalization testing with different talkers and different words.

The first generalization test consisted of the same 50 words spoken by a new talker, a talker whose tokens the subjects had not been exposed to during the experiment. The results of this generalization test are shown on the extreme right side of Figure 3 labeled GENERALIZATION. The wholistic group correctly identified 72% of the words (36 out of 50 words) whereas the dimensional group identified only 58% of the words (29 out of 50 words).

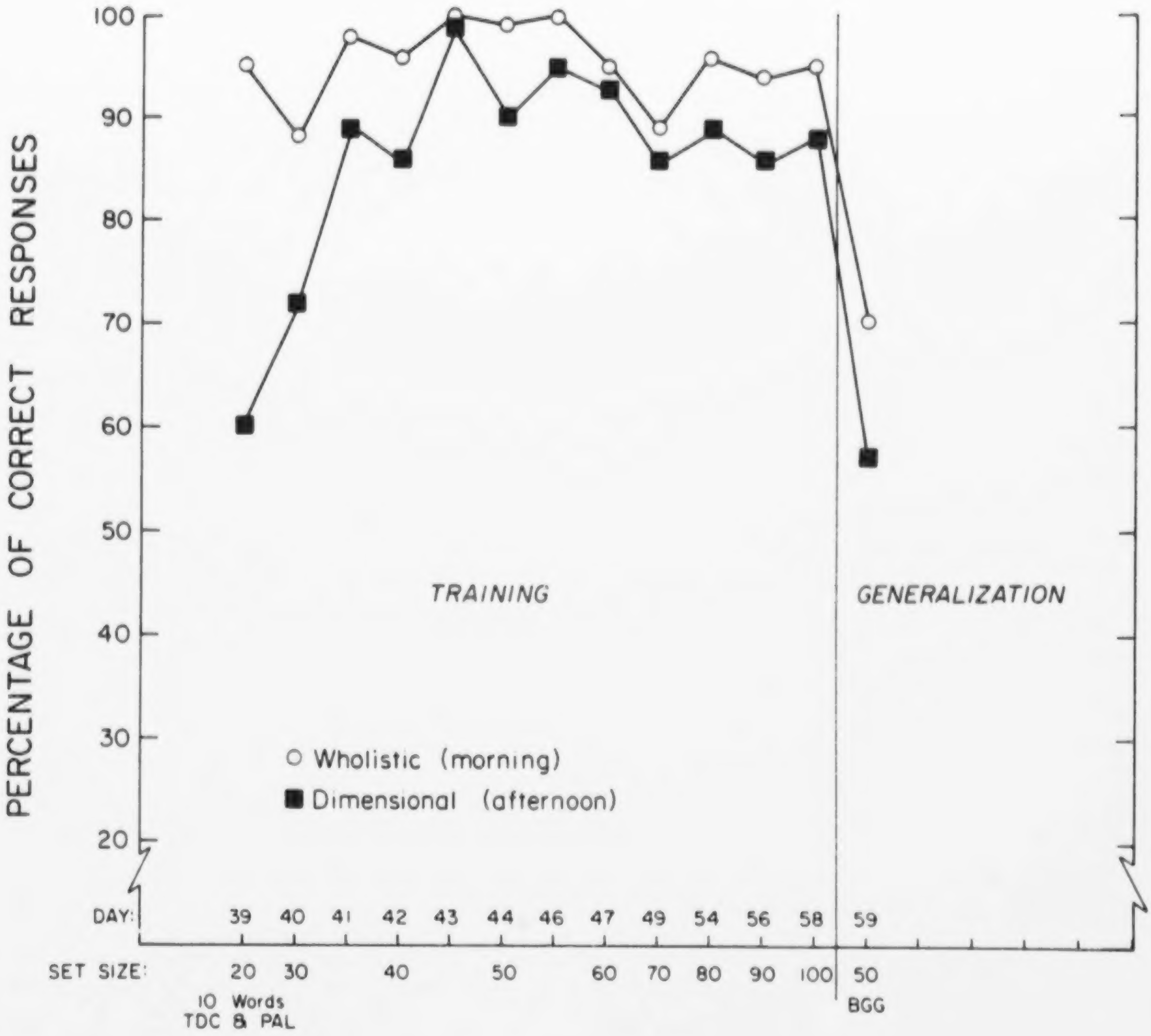


Figure 3. Daily test results graphed on a day-by-day basis are shown in the left hand portion of the figure labeled TRAINING. Results of the generalization test for the unfamiliar talker are shown in the right hand portion labeled GENERALIZATION. The set size tested, that is the number of words subjects had learned to date, is shown on the abscissa.

The second generalization test consisted of 50 additional novel words spoken by the same two talkers who produced the words used in the training set. These novel words represented the same ten dimensions and consisted of five items for each dimension, resulting in a total of 50 novel words. Subjects viewed the spectrograms in the generalization set for about ten seconds. The longer exposure time allowed subjects ample time to "analyze" each display and record their responses.

The results obtained from this generalization test with novel words showed that subjects in the wholistic group identified 15% of the words correctly while the the dimensional group identified 9.5% of the words correctly. These data reflect better generalization to unknown spectrograms than we found in our earlier study. In our earlier study, subjects correctly identified only 6% of the words.

Although performance on the novel words was low relative to the first generalization test, exact identification of the novel words was not the intended goal of this experiment. The daily test data have already shown superior learning for the wholistic group, a result that was not anticipated. From the outset, we were concerned with determining whether blocked presentation of words that shared common phonetic attributes would facilitate generalization as well as learning. Thus, the generalization results were scored by dimensions correctly identified. Instead of simply scoring a response right or wrong, we scored an item as correct if the subject indicated the correct relevant dimension for a particular dimension. For example, if "gill" or "will" was given as a response to the stimulus "kill", it was scored as correct -- the relevant dimension was correctly identified. Similarly, for the dimension /br-/, the initial consonant cluster was the only part of the response that had to be correctly identified for the response to be scored correct. Once again, the wholistic group showed better performance than the dimensional group (49% correct vs. 32.5% correct, respectively).

Finally, we examined the results separately by talker. Since the subjects learned tokens spoken by two talkers, we can examine both the exact words correctly identified and the relevant dimensions correctly identified for tokens produced by the two talkers. These results are shown in Table 2.

Insert Table 2 about here

Examination of Table 2 reveals that subjects showed better generalization performance to tokens produced by Talker 1 for both wholistic and dimensional groups. No obvious reason is apparent for this difference. Daily test scores during the course of the experiment did not yield differential scores for each talker for either group. Sometimes subjects did better on tokens spoken by Talker 1; other times they did better on tokens of Talker 2.

Table 2

Results Separately for Talkers

	TALKER 1	TALKER 2
Percentage of Novel Words Correctly Identified		
WHOLISTIC	15.33 (23 of 150)	14.66 (22 of 150)
DIMENSIONAL	10 (20 of 200)	9 (18 of 200)
Percentage of Relevant Dimensions Correctly Identified		
WHOLISTIC	54 (81 of 150)	44 (66 of 150)
DIMENSIONAL	37 (74 of 200)	28 (56 of 200)

Discussion

At the outset of this study, we expected better performance for the dimensional group because the relevant stimulus attribute(s) for each block of trials was provided explicitly for subjects. By having knowledge of the relevant dimensions, subjects should therefore be able to attend to this information to help them learn the items and, more importantly, use this information to identify new members of a category -- that is, new tokens for each relevant dimension.

Taken together, the results in this experiment show consistently better performance for the wholistic group. It appears to be that in the case of visual displays of speech, specifically speech spectrograms, explicit attention to the relevant dimensions made learning and generalization more difficult. Perhaps the overlap across stimuli that shared a dimension obscured the visual cues subjects needed to sort out the spectrograms and made them more perceptually confusable. Subjects could not learn the word-spectrogram relationship because the visual similarities outweighed the visual differences. If this argument is correct, then the performance of the wholistic group would be predicted to be better than the dimensional group because they did not (necessarily) have overlapping dimensions among stimulus items in a given study block.

Our findings also suggest that the superior performance of the wholistic group may be due to the kind of learning strategy used by these subjects. Our earlier study revealed that subjects do analyze and segment speech spectrograms and that they can describe specific parts of the spectrogram in ways that correspond to well-known acoustic-phonetic attributes. The subjects in the wholistic group may have defined certain parts of the spectrogram as their own "relevant dimension" to help them remember the specific display. When they saw another spectrogram that also included one of their own relevant dimensions, they used this information to help learn the new display. In constrast to the dimensional group, the wholistic group abstracted relevant dimensions from the stimulus items for themselves. Implicit learning may therefore have facilitated generalization while explicit presentation and ordering of the dimensions did not.

In summary, the results from the present experiment suggest that subjects learn to identify speech spectrograms more successfully when the words are presented in a mixed rather than a blocked format. Subjects trained using the wholistic approach also showed better generalization performance to a new talker and to new tokens than subjects trained using the dimensional approach. Although explicit pattern information was provided for the dimensional group, they apparently did not find this organization of speech spectrograms helpful to them during learning or generalization testing. These findings demonstrate that conscious attention to and awareness of salient visual dimensions may not necessarily facilitate perceptual learning and generalization to complex visual displays of speech as shown in spectrograms.

References

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, *58*, 955-991.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1983). Identification of speech spectrograms: Comparisons of naive versus trained observers. *Research on Speech Perception Progress Report No. 9*. Bloomington, IN: Speech Research Laboratory, Indiana University. Pp. 60-74.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1984). Recognition of speech spectrograms. *Journal of the Acoustical Society of America*, *76*, 32-43.

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

Testing the Performance of Isolated Utterance Speech Recognition Devices*

Howard C. Nusbaum, Christopher N. Davis,
David B. Pisoni, and Ella Davis

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

*This research was supported by a contract from IBM Corporation to Indiana University in Bloomington. We thank Interstate, Votan, and Texas Instruments for providing their speech recognition systems for testing purposes and for their assistance in carrying out these tests. We also thank Dr. David Pallett of the National Bureau of Standards for providing us with a copy of the speech database collected at Texas Instruments. Special thanks also go to Dr. Moshe Yuchtman for his invaluable assistance and consultation throughout many phases of this research project and to Jerry Forshee and Dave Link for their help with the instrumentation. A version of this paper appears in the Proceedings of the 1986 Voice I/O Systems Applications Conference.

Abstract

In order to choose the appropriate speech recognition device for a particular application, a number of factors must be considered carefully. These factors include performance, basic recognition capabilities (e.g., vocabulary size), price, the user/system interface, and ease of programming. Most of these factors can be assessed quite easily relative to the demands and requirements of a specific application or task. However, it is much more difficult to determine the relative performance of different speech recognition devices without carrying out systematic tests of these devices under controlled conditions. Of course, this type of comparison depends on the development of standardized methods for performance testing. At the Speech Research Laboratory, we have developed a computer-controlled system and a standardized methodology for measuring the performance of speech recognition devices under carefully controlled laboratory conditions. While this approach may not provide information about the performance of speech recognition systems in specific applications and tasks, it does permit detailed comparisons and the development of benchmarks of recognition performance under controlled conditions. By carrying out performance tests completely under computer control, we are able to eliminate errors and variability in performance introduced into testing by a human operator. Furthermore, it is possible to systematically investigate the effects of different training algorithms, vocabulary differences, talker differences, and different parameter settings.

Testing the Performance of Isolated Utterance Speech Recognition Devices

In normal conversation, a talker will issue commands, requests, and assertions by speaking directly to another person and, in general, this spoken language will be understood with a very high rate of success, even though communication by speech is not a perfect process. Communication between humans using spoken language is extremely successful because speech is highly redundant at all levels of analysis and because there are a number of linguistic and paralinguistic conventions for error correction and recovery that can be used to assist comprehension. By comparison, speaking to a speech recognition system is a much less satisfactory communication process because the talker's speaking behavior is artificially constrained to conform to the limitations of recognition technology. By comparison with human listeners, recognition performance is extremely poor. And, there are very few robust strategies for error correction. In most interactions using speech recognition devices, talkers must pause between words, they must carefully choose each word from a restricted set of alternatives, and the speech must be articulated clearly and consistently. Since the interaction between human and machine is seldom an intelligent dialogue, the type of feedback that is provided about errors may be very frustrating to the user.

Speaking to a machine using current speech technology is not as simple and effortless as talking with a person. The act of choosing words from a relatively small and restricted vocabulary (greater than 50 words, but often much less than 300 words), and of speaking in a precise and consistent fashion may require a great deal of effort and attention. Thus, the use of speech recognition systems in a variety of applications may produces human factors problems that do not arise with other interface technologies. Although some research has been directed at investigating the problems that are involved in the use of commercially available, isolated utterance, small vocabulary speech recognition devices for database retrieval, command and control, and personnel training applications (see Simpson, McCauley, Roland, Ruth, and Williges, 1985), there are still many important issues that must be addressed systematically.

One fundamental issue that has not been extensively investigated concerns the performance of speech recognition systems. Basic information regarding the performance of a speech recognition system is important for assessing the suitability of a recognition system for a particular application and can be used to determine price/performance comparisons between systems. However, there has been very little systematic research directed at investigating the performance of speech recognition systems, especially by comparison with studies investigating the performance of text-to-speech systems (e.g., Greene, Logan, & Pisoni, 1986).

Performance testing of speech recognition systems is important for several reasons. First, performance testing is an integral part of the process of improving existing recognition systems and developing new and more robust recognition algorithms. Without systematic measures of recognition performance under controlled conditions, it is almost impossible to determine whether changes in a recognition algorithm result in reliable changes in recognition accuracy. Also, detailed analyses of performance measures can provide diagnostic information about the specific problems inherent in a particular recognition algorithm and may suggest how an algorithm can be improved.

Second, performance testing provides data on speech recognition systems that may be important for determining how appropriate a particular recognition system is for a specific application. Typically, many vendors of recognition systems cite performance of their technology at 99% without specifying the conditions under which testing was carried out. Without knowing something about the structure of the vocabulary used in testing, how many talkers and tokens of speech were used, or how training was carried out, or what signal-to-noise ratio or microphone was used, it is impossible to determine what 99% correct recognition really means or how it compares to some other performance measure. To assess the relative performance of different recognition systems, it is necessary to compare performance when testing is carried out systematically under carefully controlled conditions. Performance comparisons between systems are meaningless if the testing conditions are not controlled and comparable for all the systems so that replication would be possible.

Finally, performance testing can serve an important function for advancing basic research on speech recognition. Detailed analyses of the data collected in performance tests can provide fundamental information on individual differences between talkers, the effects of noise on speech production, the role of lexical stress in distinguishing vocabulary items, as well as a number of other basic issues in speech research. Although recognition systems have not played a large role in basic speech research in the past, the potential contribution is substantial since these systems code speech produced by training and generate direct distance functions for utterances compared to the trained vocabulary. Thus, performance testing of recognition systems is important to a number of issues surrounding the development, application, and advancement of recognition technology as well as basic speech research.

In recent years, there has been increasing interest in developing standardized methods for testing speech recognition systems both to insure some degree of comparability between different tests, as well as to promote testing reliability and validity (e.g., Moore, 1986; Pallett, 1982, 1985). But despite the obvious importance of establishing standard testing methods, there is currently no concrete set of standards to govern or direct the testing of speech recognition systems. Pallett (1985) has described some general guidelines that are based, in part, on discussions arising in the IEEE Working Group on Speech I/O Systems Performance Assessment (see Baker, Pallett, & Bridle, 1983). These guidelines provide an important starting point for developing a standardized testing methodology. Pallett discusses in detail many specific issues that concern the testing process and interpretation of test results. However, beyond this initial discussion of testing issues, there has not been any attempt made to describe a standard testing procedure in sufficient detail that it could be implemented in different laboratories to permit a direct comparison of the performance of different speech recognition systems.

The need for a specific and well-described testing procedure comes from the simple observeration that no one laboratory will be able to exhaustively test all the available speech recognition systems. Furthermore, as algorithms are revised and improved, and new systems are developed, it would be a full-time effort for any single laboratory to keep up with the constant changes that occur in recognition systems. Thus, there is a definite need to distribute the process of testing recognition systems among speech research laboratories in such a way as to permit a meaningful comparison of the results from the different tests. A clearly specified and well-developed testing procedure that is adhered to carefully would insure that different laboratories would be able to carry out separate testing programs that are directly comparable. Even if particular research groups chose to go beyond the established testing methods, data collected using the standard procedure along with data from other procedures would provide a basis for comparison and interpolation. Moreover, an established testing procedure would enable a laboratory that has developed a new recognition system to test that system (using standard methods) and compare the results against earlier tests of other systems.

Part of the reason that no standard testing methodology has been specified is that there is no general agreement as to what constitutes a fair and meaningful test of the performance of a speech recognition system. Tests can be conducted using a live microphone while the talkers perform some specific task or tests can be based on a digitized database of speech. The vocabulary can be simple or difficult, small or large, and it can be based on an application or generated according to theoretical criteria (e.g., Ohala, 1982). The talkers who produce the speech for the test can represent one dialect or many. The speech can be produced under benign laboratory conditions or under more severe noise and stress conditions (e.g., Pisoni, Bernacki, Nusbaum, & Yuchtman, 1985). Training and testing can be carried out under directly comparable conditions or under conditions that are individually tailored to produce the best performance obtainable from each recognition device.

Each of these decisions regarding the conditions for testing a recognition system could have a significant impact on the performance that is measured in a given task. In general, the choices made for any particular test will reflect the specific goals of the test, as well as any constraints on the testing process. However, it is clear that there is no single test that can be carried out for a group of recognition systems that will answer all possible performance questions. Furthermore, carrying out performance tests is very time-consuming and expensive, and requires a great deal of effort and expertise. As a consequence, very few controlled performance tests of recognition systems have been reported.

In the first systematic study of speech recognition performance, Doddington and Schalk (1981) reported a comparison of seven commercially available systems ranging in cost from $500.00 to $65,000.00. These systems were tested on a database of speech collected from eight male and eight female talkers. The vocabulary consisted of the ten digits and ten control words. Ten training tokens of each vocabulary item were recorded and 16 tokens of each item were recorded for testing recognition performance. The reject threshold of each recognition system was disabled to allow more direct comparisons of recognition performance. Thus, the majority of errors were substitutions of one vocabulary item as an incorrect response to an utterance that was presented for recognition. Performance across systems ranged from .2% to 12.6% substitution errors and these errors increased, in general, with decreases in system price.

A second study reported by Baker (1982) compared the performance of the recognizers tested by Doddington and Schalk (1981) under slightly different conditions. The database used by Baker was produced by five male and five female talkers. The vocabulary consisted of the ten digits and the letter "O." Each talker produced 10,210 utterances for testing purposes and the speech was recorded simultaneously by two different microphones to compare their effects on performance. The rank-order of performance for the recognizers on both microphones generally corresponded to the results obtained by Doddington and Schalk (1981), and the magnitude of the errors was also quite similar (ranging from .2% to 16.3% for one of the microphones). Thus, the pattern of results was quite similar for two different tests conducted on two similar (but not identical databases) carried out by two different laboratories: The overall rank order of results was the same despite differences in microphones, talkers, vocabularies, and testing procedures.

Although the differences between the testing conditions in these two studies are small compared to the differences between benign laboratory enviroments and applications under more severe conditions, the general concordance between these two sets of results indicates that recognition performance is quite stable across tests. However, these studies are several years old and few, if any, of the recognizers tested in these studies are currently marketed in the same form. Most of the systems that were tested are no longer produced or marketed (e.g., Verbex 1800, Threshold Technology T-500, or Heuristics 7000). Thus, the results of these studies do not apply to currently available speech recognition devices.

Over the past year and a half, we have been carrying out a project designed specifically to investigate the performance of several commercially available speech recognition systems. The primary goal of this research has been to measure and compare the performance of speech recognition systems under controlled, laboratory conditions. In the present paper, we will only discuss the procedures and the system that we have developed for automatically testing the recognition performance of these systems.

To measure and compare the performance of isolated-utterance, small-vocabulary recognition systems, it was necessary to accomplish two other goals. The first goal was to develop a set of explicit testing procedures that could be clearly described and implemented by almost any laboratory with the necessary facilities. The second goal was to design and implement a computer-controlled testing system that would measure the performance of speech recognition devices automatically, without constant human attention.

Testing Method

Although no single test procedure can completely address all possible issues that arise in considering the assessment of recognition performance, the development of an explicit and coherent testing procedure permits the direct comparison of performance data collected in different laboratories. Furthermore, in our performance tests, we have used a standard database of speech that is currently in the public domain to provide another level of comparability between our results and the results of other tests carried out using this database.

Speech Database. The database of speech that we chose for testing purposes was collected by Doddington and Schalk (1981) and is distributed in digital and analog form by the National Bureau of Standards. This database was produced by eight male and eight female talkers and consists of two vocabularies. The TI-20 vocabulary contains the ten digits and the ten control words YES, NO, GO, START, STOP, ENTER, ERASE, RUBOUT, REPEAT, and HELP. The TI-Alpha vocabulary contains spoken versions of the 26 letters of the alphabet. For each vocabulary item, each talker produced 10 tokens for training purposes and 16 tokens for recognition testing. This database was originally digitized at 12.5 kHz with 12-bit resolution. For testing purposes, we presented the speech at 12.5 kHz, using 12 bits of resolution with a 16-bit D/A converter. The speech was low-pass filtered at 4.8 kHz.

The TI-20 vocabulary was chosen to permit comparisons of our test results with other tests carried out with this vocabulary. Also, the words in this vocabulary are extremely discriminable, with the majority of confusions occuring between GO and NO. By comparison, the alphabet is a much more difficult vocabulary because it contains several confusable subsets of letters such as the E-set. These confusable subsets differ only by a single phoneme (e.g., B and D) and therefore may engender many more substitution errors than would be produced for the TI-20.

It is important to measure performance on different vocabularies for several reasons. As noted earlier, the TI-20 vocabulary is a good choice because it has been used for other tests and therefore can serve as a standard benchmark of performance. Unfortunately, the TI-20 database has been circulating in the public domain for a while and it is always possible that recognition systems may be "tuned" specifically for this database. While it is always possible that a developer might design into a recognition system the capability of excelling on this particular database for the purposes of distorting performance assessments, it is more than likely that recognition systems will be developed, tested, and improved using this database as a benchmark in the vendors' laboratories. As a consequence, performance on this database might simply be better because algorithms were tested and improved specifically using these speech samples. By comparison however, the alphabet database has not been distributed in the public domain long enough to be used by vendors for their own development efforts. As a result, it is unlikely that any commercially available recognition system has been optimized for performance on these speech samples.

Another reason for testing performance on different databases of speech is to determine how recognition performance depends on the acoustic-phonetic structure of the vocabulary. Changes in vocabulary size and confusability are likely to have very large effects on recognition performance. By measuring performance on more than one vocabulary, it is possible to investigate the relative influence of different vocabularies on the performance of a number of different recognition systems.

Furthermore, by choosing a relatively easy vocabulary (e.g., the TI-20) and a relatively difficult vocabulary (e.g., the alphabet), it may be possible to predict the relative performance of recognition systems on an application vocabulary. If the rank ordering of recognition performance for different devices is similar for a difficult and an easy vocabulary, then the same rank ordering might hold for most vocabularies. In other words, if the same recognizers perform well on both vocabularies and other recognizers perform poorly on both vocabularies, this relative performance might be obtained for almost any vocabulary. Of course, this is an empirical question that can be answered by conducting the appropriate tests with several recognizers.

Recognition Systems. For the present study, we chose several commercially available recognition systems that are similar to each other in terms of intended use and price. Compared to the price range covered by the recognizers tested by Doddington and Schalk (1981), the recognizers we have been testing are much more similar in cost, varying from less than $1,000.00 to around $3,000.00. Furthermore, these systems are all compatible with the IBM-PC (and PC-compatible) microcomputers. The vocabulary sizes of these devices range from around 50 items to 256 items. In general, these systems are intended for use as speaker-dependent, isolated-utterance recognition devices, although some of the recognizers incorporate connected-speech algorithms. Thus, while there are several systematic differences among these systems, the overall similarity of these recognizers is much greater than in previous research. The systems that we have tested include: (1) the NEC SR-100, (2) the Interstate VocaLink CSRB, (3) the Votan VPC-2000, (4) the Dragon Systems Evaluation Board, and (5) the Texas Instruments Speech II board.

Calibration. To facilitate replication of test procedures and test results, it is important to calibrate and measure signal levels and signal/noise ratios for presenting signals to recognition systems. When a talker speaks directly to a recognition system, it is difficult to maintain accurate calibration levels. The talker must try to adjust his or her productions based on feedback from the recognition system. By comparison, with a digitized speech database, signal levels can be chosen that are optimal for testing purposes. The database collected by Doddington and Schalk (1981) was controlled so that the variation in the amplitude of words varied within a range of plus or minus 3 dB. Thus, it is possible to choose a single signal level for testing purposes using this database.

We have used two different calibration procedures for establishing an optimal signal level for testing each recognizer. In our calibration procedures, 0 dB is referenced by a 1 kHz sinewave generated at full 16-bit resolution presented at 1 mV. In the first calibration procedure, a recognizer is trained on the first five tokens of the TI-20 training set and then tested on all 16 test tokens for each of the 16 talkers in the database. One of these tests is carried out at levels varying from 2.5 dB to 12.5 dB in 2.5 dB steps. The signal level that produces the lowest error rate is used for testing the system.

We have also used a second calibration procedure in which a recognition system is trained on a single token of each of the ten digits for a talker and then tested on the exact same tokens. When trained and tested on the same tokens of speech, recognition accuracy is typically excellent, especially for the digit vocabulary. However, speech recognition systems generally return a distance measure or "similarity score" along with each recognition response. This value is a much more sensitive measure of the pattern matching process than recognition accuracy alone. Thus, the purpose of this test is to determine the signal level that produces the smallest distances or highest similarity scores for the digit vcocabulary. One test is carried out at several signal levels over a 25 dB range of amplitudes, varying from the lowest level at which all tokens are recognized in 5 dB steps. The distance or similarity scores returned by the recognizer are analyzed statistically to determine the level that produces the best performance and this signal level is then used for testing the recognizer.

Testing Procedure. For each recognition system, at least three tests are conducted for the TI-20 database and for the TI-Alpha database. Each of these three tests studies recognition performance with a different number of training tokens. In each test a recognizer is trained on one, three, or five tokens from the training set of the database, using tokens chosen in consecutive order from the beginning of the training set. A command file is prepared for each test that sets the names of the data, log, and error files on the VAX, initializes any parameters or states or data structures on the recognition system, and controls the training and testing procedure. Every recognition system is trained on the same tokens and is tested on all 16 test tokens. Thus, each test using the TI-20 generates 5120 data points and each test using the alphabet generates 6656 data points. Following the rationale of Doddington and Schalk (1981), all tests are carried out with the rejection thresholds disabled so that substitution errors are not traded off for rejection errors.

Testing performance at several levels of training provides important basic information about a recognition system. First, performance based on a single token of training indicates how a recognition system will perform under the minimal training conditions. It is important to note that training on one token is, in many respects, biased against recognition systems that are based on statistical modeling principles compared to template-based systems (Pallett, 1985). However, while this is certainly not the optimal training procedure for eliciting the highest performance from a recognition system, it is nonetheless important to understand how a system will perform under conditions of minimal training. It is highly optimistic to assume that all users of recognition systems will follow the vendors operating instructions exactly. In some cases, a user might simply try to implement a new vocabulary quickly to see how confusable it is; in other cases, users may simply try to reduce their own workload in training the system. Furthermore, while it is true that the one token used for training could be a "bad" utterance that does not represent the typical production of this vocabulary item, all recognition systems are trained on the same tokens. It is the relative performance of the recognition systems that is of greatest interest, so that even if the training token is not representative, all recognition systems will be at the same disadvantage.

Second, performance curves as a function of training can provide some indication of the optimal amount of training. The asymptote in performance will indicate how much training is required to obtain the highest recognition performance. Moreover, the performance curves will also indicate the tradeoff between increased training and increased performance. These data might suggest that the cost of additional training is not justified in terms of the size of the improved recognition performance that results.

Finally, after testing recognition systems on the TI-20 and the TI-Alpha databases after one, three, and five tokens of training, we have often carried out additional tests. The purpose of these additional tests is to comply with any specific vendor test recommendations that were not covered under the standard tests. In some cases, a recognition system may be able to use more than five training tokens, so another set of tests would be conducted under optimal training conditions for that system. In other cases, the performance of a recognition system may fall short of the vendors expectations so we carry out tests to attempt to diagnose more carefully the performance of the system and to try to improve performance by modifying the test conditions systematically. These auxilliary tests provided additional information about how performance might be improved when using a recognition system and, by comparison to our standard series of tests, they may indicate how much improvement might be expected.

Automatic Testing of Recognition Performance

At the present time, almost all testing of speech recognition systems is carried out under manual control by a human operator. However, in order to collect a large number of recognition test trials, control by a human operator is simply too costly. If tests are carried out automatically under computer control, data can be collected 24 hours a day for seven days a week, making the testing process much more efficient and reducing costs and possible human errors in testing.

The basic concept for this computer-controlled testing system derives from the paradigm of testing human subjects in speech perception experiments. In these experiments, human subjects are tested under real-time control of a computer that presents speech signals over headphones and collects the subjects' responses to these stimuli. Speech signals are stored in digital waveform files on a large disk and are retrieved on demand and converted to analog form for presentation to subjects. Subjects respond by pressing buttons on a computer-controlled response box, or by pressing keys on a terminal keyboard. The concept is quite straightforward: The computer controls the testing session by presenting speech signals and recording responses.

This is the paradigm we chose to employ for testing speech recognition systems. A computer controls the entire testing procedure, and a recognizer serves as the experimental subject. The controller presents speech signals to the recognizer and the recognizer generates a response which is sent back to the controller. Carrying out this testing process automatically accomplishes several objectives. First, the environment and testing conditions can be carefully controlled. This makes it possible to systematically investigate a number of factors that may have a significant impact on recognition performance. By storing speech in large digital databases, changing testing conditions such as talker gender or effects of environmental noise only requires specifing a different database without changing any other testing conditions. Second, responses are collected by the computer thereby eliminating possible human error and experimenter bias in data collection. Moreover, this permits rapid analysis and re-analysis of data using a variety of different scoring procedures. Third, by automating the testing procedure, it becomes easier to parametrically investigate changes in training protocols or parameter settings that would ordinarily be too expensive in personnel costs if carried out by human operators using the conventional approach.

There are several important capabilities that are needed to automate testing of speech recognition systems. First, at some level, a description of the functional operation of a recognizer is necessary. That is, the controller must be able to initiate each function of the recognizer that is needed to carry out testing. Second, the controller must be able to communicate with the recognition system to convey commands and receive responses. Third, the controller must be able to access databases of speech and present these signals to the recognizer according to some experimenter-defined testing protocol. Finally, the controller must maintain complete descriptions of the testing procedure, errors encountered during testing, and the recognition responses that are generated as data.

We have implemented these basic capabilities by dividing the control system into two parts: (1) a virtual device controller and interface (VDC) for speech recognition systems, and (2) a device dependent interface (DDI) that is specifically programmed for each individual recognizer. The virtual device controller and interface embodies a model of a generic speech recognition system. This model includes several functions and parameters that can be manipulated by a researcher. The training and testing protocol for each experiment is described using a command language to program the generic recognition model in the controller. The VDC communicates these commands to the DDI which translates the generic recognizer commands into the specific commands and syntax of the recognizer that is being tested. Thus, the DDI serves as the communications host and translator for the recognition system that is being tested.

Commands and data are passed between the VDC and the DDI using a communications protocol to insure proper handshaking and the integrity of information. In addition, the VDC coordinates the events of the testing procedure and presents speech to the recognizer and stores data and errors in disk files along with a log of the communications between the VDC and DDI. We call this system SPERTES (SPEech Recognition TEsting System).

Implementation of SPERTES

SPERTES instantiates the VDC and DDI as C programs running on a high speed minicomputer and a microcomputer. A VAX-11/750 serves as the controller presenting speech from a digital waveform database stored on RA-81 disks. The speech is converted into analog form using a DSC-200 16-bit digital-to-analog converter with a DSC-240 amplifier under the control of the VDC. An IBM-PC serves as the physical host for a speech recognition system. Commands to the recognizer are sent over an RS-232 serial line to the PC from the VAX. Responses from the recognizer are sent from the PC back to the VAX over this serial line.

The virtual device controller and interface has been implemented in C on the VAX-11/750 under the VMS operating system. For each recognition system, a new device dependent interface is written in C and it runs on the IBM-PC under PC-DOS. Communications between the VAX and the PC take place over the serial line at 9600 baud.

To test a recognition system, the recognizer must first be physically interfaced to the host IBM-PC, either over a serial line or in a bus slot. Then a device dependent interface is written that translates generic recognizer commands sent by the VDC into the format necessary to control the functions of the specific recognizer. Also the DDI translates the responses of the recognition system (e.g., error messages or recognition results) into the generic format expected by the VDC. Thus, the DDI manages communications between the PC and the recognizer, sending commands to the recognizer and collecting responses. In addition, the DDI communicates with the VDC running on the VAX, receiving commands and sending back responses.

Once a recognizer is interfaced to the host, and a DDI is developed and tested, the recognizer can be controlled by command files read on the VAX by the VDC software. A command file is written to control each testing session. In general, there are three classes of commands that can be used in a command file: (1) "set local" commands control some aspect of the VDC software, (2) "relay" commands allow access to recognizer-specific commands that do not have generic counterparts in the VDC, and (3) recognition commands control training and testing.

The "set local" commands control the operation of the VDC and do not affect operation of the DDI or the recognition system. One function of these commands is to specify the names of data files, log files, and error files on the VAX. Another function is to select which analog output line will be used to present speech to the recognition device. The current configuration of the DSC-200 on the Speech Research Laboratory VAX includes four analog output lines allowing up to four recognition tests to run concurrently. A "set local" command is used to initiate a debugging mode that presents the researcher with detailed information about the operation of the VDC. Also using this command, it is possible to set a delay between VAX-to-PC communications and speech output to accomodate differences in response time for different recognizers. Finally,"set local" can be used to change the error limit which specifies the number of times a command will be tried after an error is encountered.

The "relay" commands permit direct access to the functions of a recognizer that may not be part of the generic recognizer model, but may be important to conducting a test, such as gain control or recognizer-specific instructions or parameters. Another feature that may be invoked using "relay" is separate indexing of templates in a vocabulary. Normally, each vocabulary item is "enrolled" using a training token to create one template for each vocabulary item. Subsequent tokens of each vocabulary item then "update" or "retrain" the initial template creating a representation for each vocabulary item that contains more information about the pattern of a word than is created after enrollment on a single token. In some recognition systems, the updated representation may be a single template that incorporates pattern information from several tokens. On other systems, the updated representation may actually be a cluster of individual templates that are created one for each training pass and then are internally mapped onto a single vocabulary item. The "relay" command that instantiates separate indexing of templates in the DDI (called "overloading" in SPERTES), implements this latter form of updating in the DDI software. With overloading, each repeated training pass with a vocabulary item creates a new template in the recognition system that is treated by the recognition system as a new vocabulary item rather than an update of an old vocabulary item. These separate templates in the recognizer are mapped in the DDI onto single vocabulary items and returned as such to the VDC. Thus, separate indexing through the overloading feature causes the DDI to maintain a table of template numbers for each vocabulary item so that it is possible to train a recognizer on more than one token of a word and still have access to the individual templates. Overloading is not used as part of the normal performance measurement testing procedure, because it may penalize any recognition system that optimizes its models or templates for vocabulary items based on statistical properties of the vocabulary. However, overloading has proved a useful tool for investigating the operation of updating algorithms. By using overloading it is possible to determine the distribution of recognition responses over the different tokens stored separately as templates.

Finally, recognition commands control training, retraining or updating, and recognition testing. Training refers to the initial "enrollment" of a template or model in the recognizer for a word. Retraining is used for recognizers that allow multiple training passes to "enrich" an initially enrolled template or model. Recognition refers to the testing phase in which speech is sent to the recognizer and responses are collected from it.

The flags and parameter settings for these commands indicate to the VDC which words or groups of words are to be used in training or testing, and which talkers and tokens are to be accessed from the speech database that is on-line on disk. The parameter "full" accesses all the tokens of a particular word for a given talker, while "random" selects one token at random, and "s#e#" uses the tokens in numerical order from the starting number "s#" to the ending number "e#" (where the # stands for a token number). In addition, vocabulary subsets are defined such as "ti20" for the Doddington and Schalk (1981) vocabulary, "digits" for the digit set, "alpha" for the alphabet, "eset" for the E-set of the alphabet (i.e., B,D,G,P,T,C,Z,E,V). Also, any word in the vocabulary can be accessed by reference to a symbol such as "HELP." One form of the command allows the researcher to talk to the recognizer instead of using the database. Finally, it is possible to specify any valid waveform file name using full VMS pathname conventions for training, retraining, or recognition.

In order to perform a test of a recognition system, a command file is written that describes the experimental protocol. This command file uses the SPERTES commands to specify the names of files associated with the test, set parameters for the recognition test, train the recognizer on specific utterances, and finally present utterances for recognition and record responses. Using the SPERTES command set, a large number of experiments are possible. A recognition system can be trained and tested on the speech of one talker at a time, or trained on one talker and tested on several talkers, or any other combination of training and testing protocols using a stored database of speech.

For each testing session, the VDC creates several different files as output. A log file is created that contains a complete record of all commands sent from the VDC to the DDI and all the responses back from the DDI, and also includes all the local VDC actions. A second file is created containing any error messages that were generated by the VDC or the DDI during a test. This file can be used to determine if any problems occurred during a test and can aid in the diagnosis of those problems. Finally, the VDC creates a data file that contains all responses produced by the DDI from the recognition system during recognition testing. One record is stored in this file for each utterance presented for recognition. Each record of this file consists of the word that was presented to the recognition system, the first candidate returned by the system and its distance or similarity score, followed by any other candidates generated during the recognition trial. Data files are analyzed to determine substitution error rates, rejection error rates, confusion matrices, and inter-vocabulary distances.

Discussion

In summary, we have developed a software system called SPERTES that automatically controls the testing of speech recognition systems. SPERTES consists of a virtual device controller that implements a generic model of a speech recognition device and a device dependent interface that translates messages and commands between the generic model and recognizer-specific format. This automated testing system for speech recognition devices reduces human error and bias in the measurement of recognition performance and allows precise manipulation of signal and testing conditions. Moreover, the command set used by SPERTES allows a great deal of flexibility in defining test protocols and recognition experiments.

In addition, we have developed an explicit testing procedure that provides directly comparable performance measures for different recognition systems. Following a calibration procedure to determine the optimal signal level for presenting speech, each recognition system is tested on the TI-20 and alphabet vocabularies with one, three, and five training tokens. This test provides information about the performance of systems that receive minimal training and about the improvements in performance that can be expected to occur for each system with increased training. Testing on an easy vocabulary and difficult vocabulary should span a wide range of possible performance levels and thus allow performance comparisons between systems as a function of confusability of the vocabulary.

One important issue that arises in considering the measurement of recognition performance concerns the ability to predict actual performance of a specific recognition system in a particular application based on benchmark laboratory tests. If laboratory results bear no relationship to recognition performance obtained in an application environment, then the utility of benchmark testing is problematic. Indeed, there is some question as to whether performance on benchmark tests carried out under controlled laboratory conditions are predictive of performance of recognition systems under application conditions (Pallett, 1985). However it should be remembered that without data, any kind of prediction is impossible. In addition, recognition systems are physical devices and the performance of these systems is not arbitrary. Thus, it is quite unlikely that performance data obtained in controlled laboratory tests will have no predictive validity at all. While it is probably true that the best predictor of performance in a specific application is a test that was carried out completely simulating all the conditions of the application, the important issue is to determine precisely *how* to predict performance in specific applications based on laboratory data.

Let us take an example. If the recognition vocabulary, application environment, user population, and user interface are all held constant, the major variable affecting performance will be differences among the recognition systems (see Nusbaum & Pisoni, 1986). Of course, it is very unlikely that the absolute level of recognition performance obtained in a benign laboratory benchmark test will also be *identical* to performance obtained in any real application under more severe field conditions. However, it is not always necessary to predict the *absolute* performance of a recognizer in an application, but only the *relative* performance of different systems for that application. In other words, the primary issue is to predict which system will perform best in an application, and not to determine exactly how accurate recognition will be for each system in the application.

If the relative performance of different recognition systems were invariant over conditions, direct extrapolation from benchmark tests to relative performance in an application would be straightforward. But it is unlikely that the rank ordering of systems based on performance will be constant across all conditions. Therefore, the goal of laboratory testing should be to provide not only a standard basis for comparing different recognizers, but also to establish a range of performance for each recognition system to permit comparisons of performance distributions. In future research, it will be important to determine precisely how predictive laboratory performance data are of applications performance and to develop laboratory tests that are specifically designed for the purpose of prediction.

References

Baker, J. M. (1982). The performing arts -- how to measure up. In D. S. Pallett (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology*. Gaithersberg, MD: National Bureau of Standards.

Baker, J. M., Pallett, D. S., & Bridle, J. S. (1983). Speech recognition performance assessment and available databases. In *Proceedings of ICASSP-83*. New York: IEEE.

Doddington, G. R., & Schalk, T. B. (1981). Speech recognition: Turning theory to practice. *IEEE Spectrum*, *18*, 26-32.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers*, *18*, 100-107.

Moore, R. K. (1986). The NATO research study group on speech processing: RSG10. In *The Official Proceedings of Speech Tech'86: Voice Input/Output Applications Show and Conference*. New York: Media Dimensions.

Nusbaum, H. C., & Pisoni, D. B. (1986). Human factors issues for the next generation of speech recognition systems. In *The Official Proceedings of Speech Tech'86: Voice Input/Output Applications Show and Conference*. New York: Media Dimensions.

Ohala, J. (1982). Calibrated vocabularies. In D. S. Pallett (Ed.), *Proceedings of the Workshop on Standardization for Speech I/O Technology*. Gaithersberg, MD: National Bureau of Standards.

Pallett, D. S., Ed. (1982). *Proceedings of the Workshop on Standardization for Speech I/O Technology*. Gaithersberg, MD: National Bureau of Standards.

Pallett, D. S. (1985). Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards*, *90*, 371-387.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., & Yuchtman, M. (1985). Some acoustic-phonetic correlates of speech produced in noise. In *Proceedings of ICASSP 85*. New York: IEEE Press.

Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., & Williges, B. H. (1985). System design for speech recognition and generation. *Human Factors*, *27*, 115-141.

IV. PUBLICATIONS

Charles-Luce, J. Comparison in Bambara: An infinitival verb phrase. *Studies in African Linguistics*, *17*, 199-212.

Charles-Luce, J. Word final devoicing in German and the effects of phonetic and sentential contexts. *Journal of Phonetics*, 1985, *13*, 309-324.

Elbert, M. and Gierut, J. A. *Handbook of Clinical Phonology: Approaches to Assessment and Treatment*. San Diego: College Hill Press, 1986.

Gierut, J. A. Sound change: A phonemic split in a misarticulating child. *Applied Psycholinguistics*, 1986, *7*, 57-68.

Gierut, J. A., and Dinnsen, D. A. On word-initial voicing: Converging sources of evidence in phonologically disordered speech. *Language and Speech*, 1986, *29*, 97-114.

Greene, B. G. Perception of synthetic speech by nonnative speakers of English. *Proceedings of the Human Factors Society, Volume 2*. Santa Monica, CA: Human Factors Society, 1986. Pp. 1340-1343.

Greene, B. G., Logan, J. S., and Pisoni, D. B. Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, and Computers*, 1986, *18*, 100-107.

Greenspan, S. L. Semantic flexibility and referential specificity of concrete nouns. *Journal of Memory and Language*, 1986, *25*, 539-557.

Greenspan, S. L., Nusbaum, H. C. and Pisoni, D. B. Perception of synthetic speech: Some effects of training and attentional limitations. *Proceedings of the 5th Voice Input/Output Applications Conference*. Sunnyvale, CA: Lockheed, 1985.

Kewley-Port, D. Converging approaches towards establishing invariant acoustic-phonetic cues. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986. Pp. 193-197.

Kubaska, C. A. and Aslin, R. N. Categorization and normalization of vowels by 3-year-old children. *Perception & Psychophysics*, 1985, *37*, 355-362.

Luce, P. A. A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, 1986, *39*, 155-158.

Luce, P. A. and Charles-Luce, J. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, 1985, *78*, 1949-1957.

Nusbaum, H. C. Human factors considerations in the design of large vocabulary speech recognition devices. *Proceedings of Speech Tech '86*. NY: Media Dimensions, 1986. Pp. 139.

Nusbaum, H. C., Davis, C. K., Pisoni, D. B., and Davis, E. Testing the performance of isolated utterance speech recognition devices. *Proceedings of the Voice Data Entry Systems Applications Conference*. Palo Alto: AVIOS, 1986. Pp. 393-408.

Nusbaum, H. C. and Pisoni, D. B. Human factors issues for the next generation of speech recognition systems. *Proceedings of Speech Tech '86*. NY: Media Dimensions, 1986. Pp. 140-144.

Nusbaum, H. C. and Pisoni, D. B. The role of structural constraints in auditory word recognition. *Montreal Symposium on Speech Recognition*, 1986. Pp. 57-58.

Pisoni, D. B. A brief overview of speech synthesis and recognition technologies. *Proceedings of the Human Factors Society, Volume 2*. Santa Monica, CA: Human Factors Society, 1986. Pp. 1326-1330.

Pisoni, D. B. Contextual variability and the problem of acoustic-phonetic invariance in speech. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986. Pp. 154-161.

Pisoni, D. B. and Luce, P. A. Speech perception: Research, theory, and the principal issues. In E. C. Schwab and H. C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines: Speech Perception*. Volume 1. New York: Academic Press, 1986. Pp. 1-50.

Pisoni, D. B. and Nusbaum, H. C. Developing methods for assessing the performance of speech synthesis and recognition systems. *Proceedings of the Human Factors Society, Volume 2*. Santa Monica, CA: Human Factors Society, 1986. Pp. 1344-1348.

Pisoni, D. B., Nusbaum, H. C. and Greene, B. G. Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 1985, *73*, 1665-1676.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A. and Slowiaczek, L. M. Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, 1985, *4*, 75-95.

Pisoni, D. B., Yuchtman, M. and Hathaway, S. N. Effects of alcohol on the acoustic properties of speech. In *Alcohol, Accidents, and Injuries*. Warrendale, PA: Society of Automotive Engineers, 1986. Pp. 131-150.

Salasoo, A. Cognitive processing in oral and silent reading comprehension. *Reading Research Quarterly*, 1986, *21*, 59-69.

Schwab, E. C. and Nusbaum, H. C. (Eds.), *Pattern Recognition by Humans and Machines: Volume 1, Speech Perception*. NY: Academic Press, 1986.

Schwab, E. C. and Nusbaum, H. C. (Eds.), *Pattern Recognition by Humans and Machines: Volume 2, Visual Perception*. NY: Academic Press, 1986.

Schwab, E. C., Nusbaum, H. C. and Pisoni, D. B. Some effects of training on the perception of synthetic speech. *Human Factors*, 1985, *27*, 395-408.

Slowiaczek, L. M. and Dinnsen, D. A. Neutralization and word final devoicing in Polish. *Journal of Phonetics*, 1985, *13*, 325-341.

Slowiaczek, L. M. and Nusbaum, H. C. Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, 1985, *27*, 701-712.

Slowiaczek, L. M. and Pisoni, D. B. Effects of phonological similarity on priming in auditory lexical decision. *Memory & Cognition*, 1986, *14*, 230-237.

Stemberger, J. P. and MacWhinney, B. Form-oriented inflectional errors in language processing. *Cognitive Psychology*, 1986, *18*, 329-354.

Stemberger, J. P. and MacWhinney, B. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition*, 1986, *14*, 17-26.

Stemberger, J. P and Treiman, R. The internal structure of word-initial consonant clusters. *Journal of Memory and Language*, 1986, *25*, 163-180.

Walley, A. C., Smith, L. B. and Jusczyk, P. W. The role of phonemes and syllables in the perceived similarity of speech sounds for children. *Memory & Cognition*, 1986, *14*, 220-229.

Yuchtman, M. and Nusbaum, H. C. Using template structure information to improve speech recognition performance. *Proceedings of the Voice Data Entry Systems Applications Conference*. Palo Alto: AVIOS, 1986. Pp. 375-391.

Papers Accepted for Publication (In Press):

Connine, C. M. and Clifton, C., Jr. Interactive uses of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, (in press).

Gierut, J. A. On the assessment of productive phonological knowledge. *National Student Speech, Language, Hearing Association Journal*, (in press).

Gierut, J. A., and Dinnsen, D. A. On predicting ease of phonological learning. *Applied Linguistics*, (in press).

Gierut, J. A., Elbert, M., and Dinnsen, D. A. A functional analysis of phonological knowledge and generalization learning in misarticulating children. *Journal of Speech and Hearing Research*, (in press).

Greene, B. G. and Pisoni, D. B. Perception of synthetic speech by adults and children: Research on processing voice output from text-to-speech systems. In L. E. Bernstein (Ed.), *The Vocally Impaired: Volume II Basic Research and Technology*. New York: Academic Press, (in press).

Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition*, (in press).

Luce, P. A. Similarity neighborhoods and word frequency effects in visual word identification: Sources of facilitation and inhibition. *Journal of Memory and Language*, (in press).

Luce, P. A. and Pisoni, D. B. Speech perception: New directions in research, theory, and application. In H. Winitz (Ed.), *Human Communication and Its Disorders*. Norwood, NJ: Ablex, (in press).

Pisoni, D. B. Auditory perception of complex sounds: Comparisons of speech vs. nonspeech signals. In W. A. Yost and C. S. Watson (Eds.), *Complex Sound Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, (in press).

Pisoni, D. B. Some measures of intelligibility and comprehension. In J. Allen, D. H. Klatt, and S. Hunnicutt (Eds.), *From Text to Speech: The MITalk System*. Cambridge, UK: Cambridge University Press, (in press).

Pisoni, D. B. and Luce, P. A. Acoustic-phonetic representations in the mental lexicon. *Cognition*, (in press).

Pisoni, D. B. and Luce, P. A. Trading relations, acoustic cue integration, and context effects in speech perception. In M. E. H. Schouten (Ed.), *Proceedings of the NATO Advance Research Workshop on Psychophysics and Speech Perception*, Utrecht, 1986, (in press).

Slowiaczek, L. M. and Pisoni, D. B. Speech perception. In *McGraw-Hill Encyclopedia of Science and Technology*, 4th edition, (in press).

Slowiaczek, L. M., Nusbaum, H. C. and Pisoni, D. B. Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition*, (in press).

Stemberger, J. P. Between-word processes in child phonology. *Journal of Child Language*, (in press).

Stemberger, J. P. and Lewis, M. Reduplication in Ewe: Accommodation to phonological errors. *Phonology Yearbook, No. 3*, (in press).

V. Speech Research Laboratory Staff, Associated Faculty, and Technical Personnel

(9/1/85 - 12/31/86)

Research Personnel:

David B. Pisoni, Ph.D. --------------- Professor of Psychology and Director
Beth G. Greene, Ph.D. ---------------- Research Scientist and Associate Director

Daniel A. Dinnsen, Ph.D. ------------- Professor of Linguistics
Howard C. Nusbaum, Ph.D. ------------- Assistant Research Scientist*
Steven L. Greenspan, Ph.D. ----------- Research Associate**
Moshe Yuchtman, Ph.D. ---------------- Research Associate

Cynthia M. Connine, Ph.D. ------------ NIMH Post-doctoral Fellow
Judith A. Gierut, Ph.D. -------------- NIH Post-doctoral Fellow
John W. Mullennix, Ph.D. ------------- NIH Post-doctoral Fellow
Van Summers, Ph.D. ------------------- NIH Post-doctoral Fellow
Kazunori Ozawa, B.E.E. --------------- Visiting Scientist***
Jan Charles-Luce, M.A. --------------- NIH Pre-doctoral Fellow
John Kuster, B.A. ------------------ Graduate Research Assistant
John S. Logan, B.S. ------------------ Graduate Research Assistant
Paul A. Luce, B.A. ------------------- Graduate Research Assistant
Laura M. Manous, B.A. ---------------- Graduate Research Assistant
Christopher S. Martin, B.A. ---------- Graduate Research Assistant
Robert I. Pedlow, M.Sc. -------------- Graduate Research Assistant
Louisa M. Slowiaczek, B.A. ----------- Graduate Research Assistant+

Technical Support Personnel:

Christopher K. Davis ----------------- Applications Programmer
Ella Davis --------------------------- Programmer
Michael J. Dedina, B.A. -------------- Research Assistant/Programmer
Jerry C. Forshee, M.A. --------------- Computer Systems Analyst
David A. Link ------------------------ Electronics Engineer
Gary Link ---------------------------- Technical Assistant
Mary F. Stapleton -------------------- Administrative Secretary

Kimberly Baals ----------------------- Undergraduate Research Assistant
Lisa Huber --------------------------- Undergraduate Research Assistant
Amy Lawlor --------------------------- Undergraduate Research Assistant
Penny Mechley ------------------------ Undergraduate Research Assistant
Michael Stokes ----------------------- Undergraduate Research Assistant

---

*Now at the University of Chicago, Chicago, IL
**Now at AT&T Bell Labs, Naperville, IL
***Research Engineer, C&C Information Technology Research Laboratories, NEC Corporation, Kawasaki, Japan
+Now at Loyola University of Chicago, Chicago, IL